CONTRIBUTIONS TO COMPUTER EXPERIMENTS AND BINARY TIME SERIES

A Thesis
Presented to
The Academic Faculty

by

Ying Hung

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
August 2008

CONTRIBUTIONS TO COMPUTER EXPERIMENTS AND BINARY TIME SERIES

Approved by:

Dr. C. F. Jeff Wu, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Roshan Joseph Vengazhiyil, Co-advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Kwok L. Tsui
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Ming Yuan
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Shreyes Melkote
School of Mechanical Engineering
Georgia Institute of Technology

Date Approved: 6 May 2008

To my parents,

for their support, inspiration and encouragement

during this challenging journey.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my appreciation to all who have influenced, stimulated, expedited, and warmly supported my work in various ways.

First and foremost, I would like to express my deep and sincere gratitude to my advisor, Professor C. F. Jeff Wu, for his guidance, assistance, encouragement, and hearty support at all phases of my doctoral program. He took care of me both academically and personally in every conceivable way. He has not only been my academic advisor, but also a great mentor for my graduate student life.

I am extremely thankful to my co-advisor, Professor Roshan Joseph Vengazhiyil, for his guidance on my research topic and overwhelming support during my studies. His constant inspiration helped me overcome many problems and achieve this milestone.

I would like to thank Dr. Shreyes Melkote and Dr. Cheng Zhu for having extended their support to my research and having guided and inspired me in various ways. I am also thankful to Dr. Kwok Leung Tsui and Dr. Ming Yuan for serving on my dissertation committee and for their valuable comments and suggestions.

I am very thankful to my lab members Dr. Tirthankar Dasgupta, Dr. Abhyuday Mandal, Dr. Zhiguang Qian, Xinwei Deng, Lulu Kang, Nagesh Adiga, and Huizhi Xie who shared time, space and knowledge with me at Georgia Tech. I consider myself very fortunate to be able to spend a lot of time with these outstanding people.

Last, but by no means the least, my heartfelt appreciation and gratitude goes to my family, especially my parents and Ang Lee, for their constant support and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This thesis consists of two parts. The first part focuses on design and analysis for computer experiments and the second part deals with binary time series and its application to kinetic studies in micropipette experiments.

The first part of the thesis contains three chapters. In the first chapter, a new experimental design for computer experiments is developed. The second chapter proposes a new analysis method for computer experiments. A novel methodology for the design and analysis of experiments with nested factors is developed in the third chapter. The second part of the thesis is included in chapter four, where a new multiple binary time series model and related inference are developed.

The research described in chapter one is concerned with optimal design for computer experiments. Because deterministic models are used for experiments, the output of a computer experiment (or code) is not subject to random variations, which makes the design of computer experiments different from that of physical experiments. Latin hypercube designs (LHDs) have been used extensively in the computer experiments literature. A randomly generated LHD can have a systematic pattern: the variables may be highly correlated or the design may not have good space-filling properties. There are procedures to find good LHDs by minimizing the pairwise correlations or maximizing the inter-site distances. In this chapter, it is shown that these two criteria need not agree with each other. In fact, maximization of inter-site distances can result in LHDs where the variables are highly correlated and vice versa. Therefore, a multi-objective optimization approach is proposed to find good LHDs by combining

correlation and distance performance measures. A new exchange algorithm for efficiently generating such designs is also proposed. Several examples are presented to show that the new algorithm is fast and that the obtained designs are good in terms of both the correlation and distance criteria.

In computer experiments, kriging is a useful method for developing metamodels for product design optimization. The most popular kriging method, known as ordinary kriging, uses a constant mean in the model. In the second chapter, a modified kriging method is proposed, which has an unknown mean model. Therefore it is called blind kriging. The unknown mean model is identified from experimental data using a Bayesian variable selection technique. Many examples are presented which show remarkable improvement in prediction using blind kriging over ordinary kriging. Moreover, the blind kriging predictor is easier to interpret and more robust to misspecification in the correlation parameters.

The third chapter addresses problems related to computer experiments with branching and nested factors. In many experiments, some of the factors exist only within the level of another factor. Such factors are often called nested factors. A factor within which other factors are nested is called a branching factor. For example, suppose we want to experiment with two processing methods. The factors involved in these two methods can be different. Thus, in this experiment the processing method is a branching factor and the other factors are nested within the branching factor. Design and analysis of experiments with branching and nested factors are challenging and have not received much attention in the literature. Motivated by a computer experiment in a machining process, we develop optimal Latin hypercube designs and kriging methods that can accommodate branching and nested factors. Through the application of the proposed methods, optimal machining conditions and tool edge geometry are attained, which resulted in a remarkable improvement in the machining process.

The fourth chapter deals with binary time series analysis with application to cell adhesion frequency experiments. Repeated adhesion frequency assay is the only published method for measuring the kinetic rates of cell adhesion. Cell adhesion plays an important role in many physiological and pathological processes. Traditional analysis of adhesion frequency experiments assumes that the adhesion test cycles are independent Bernoulli trials. This assumption can often be violated in practice. Motivated by the analysis of repeated adhesion tests, a binary time series model incorporating random effects is developed in this chapter. A goodness-of-fit statistic is introduced to assess the adequacy of distribution assumptions on the dependent binary data with random effects. The asymptotic distribution of the goodness-of-fit statistic is derived and its finite-sample performance is examined via a simulation study. Application of the proposed methodology to real data from a T-cell experiment reveals some interesting information, including the dependency between repeated adhesion tests. These results provide some quantitative evidence to the speculation that cells can have "memory" in their adhesion behavior.

CHAPTER I

ORTHOGONAL-MAXIMIN LATIN HYPERCUBE DESIGNS

[1]

1.1 Introduction

Computer experiments are widely used for the design and development of products (for examples, see Fang, Li, and Sudjianto 2006). In computer experiments, instead of physically doing an experiment on the product, mathematical models describing the performance of the product are developed using engineering/physics laws and solved on computers through numerical methods such as the finite element method. Because deterministic models are used for experiments, the output of a computer experiment is not subject to random variations, which makes the design of computer experiments different from that of physical experiments (see Sacks et al. 1989). For example, replication is not required. In fact, it is desirable to avoid replicates when projecting the design on to a subset of factors. This is because a few out of the numerous factors in the system usually dominate the performance of the product (known as effect sparsity principle). Thus a good model can be fitted using only these few important factors. Therefore, when we project the design on to these factors, replication is not required. This can be achieved by using a Latin Hypercube Design (LHD) (McKay, Beckman, and Conover 1979). An LHD has the property that by projecting an n-point design on to any factor, we will get n different levels for that factor. This property makes an LHD very suitable for computer experimentation.

[1]The paper based on this chapter appeared in *Statistica Sinica*, **18**, 171-186, 2008.

Suppose the n levels of a factor are denoted by $1, 2, \cdots, n$. Figure 1a shows an LHD with two factors in six points. In general, an n-run LHD can be generated using a random permutation of $\{1, 2, \cdots, n\}$ for each factor. Each permutation leads to a different LHD. For k factors, we can thus obtain $(n!)^k$ LHDs. Figure 1b shows one such LHD. Clearly, this is not a good design. It is not good due to the following two reasons. First, the two factors are perfectly correlated. Therefore, we will not be able to distinguish between the effects of the two factors based on this experiment. Second, there is a large area in the experimental region that is not explored. Therefore, if we use such a design to develop a prediction model, then the prediction will be poor in those unexplored areas.

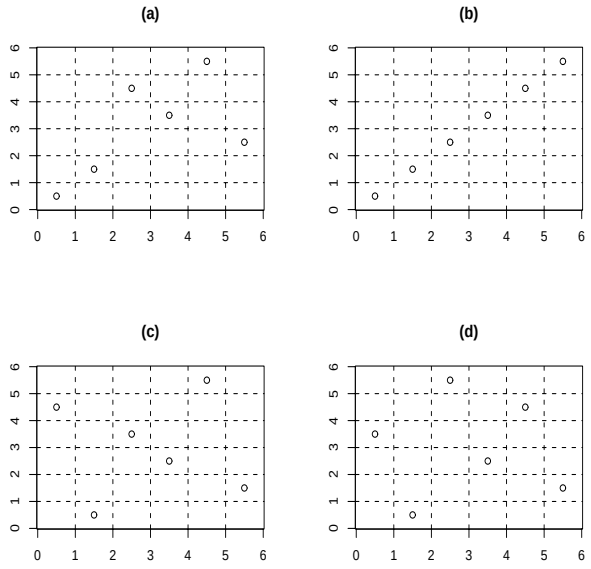

Figure 1: LHDs

There has been some work in the literature to avoid the above problems and obtain a "good" LHD. The idea is to find the best design by optimizing a criterion that describes a desirable property of the design. Iman and Conover (1982), Owen (1994), and Tang (1998) proposed to find designs minimizing correlations among factors. Figure 1c shows the optimal LHD found by the procedure in Tang (1998),

which is clearly much better than the one in Figure 1a and 1b. As discussed before, apart from the correlations we are also interested in spreading the points out across the experimental region. This is the idea behind space-filling designs. Morris and Mitchell (1995) proposed to find the best LHD by maximizing the minimum distance between the points. The optimal LHD under this criterion is shown in Figure 1d. Other approaches to find good LHDs are given by Owen (1992), Tang (1993), Park (1994), Ye (1998), Ye, Li, and Sudjianto (2000), and Jin, Chen, and Sudjianto (2005).

The minimum pairwise correlation between the factors and the maximum distance between the points are both good criteria for finding optimal LHDs. Intuitively, minimizing correlation should spread out the points and maximizing the distance between the points should reduce the correlation. But in reality, there is no one-to-one relationship between these two criteria. In fact, the designs obtained by these two criteria can be entirely different. To illustrate this, consider again an LHD with six points and two factors. There are a total of $(6!)^2 = 518,400$ LHDs. The designs can be ranked based on the maximin distance criterion (Mitchell and Morris 1995), where the rank 1 is given to the best design. They are plotted in Figure 2 against absolute values of correlations (there are a total of 113 different combinations of correlations and maximin ranks in this example). We can see that the points are highly scattered showing that the minimization of one criterion may not lead to the minimization of the other criterion (see Figure 3 for an example.) The problem becomes more serious as the number of points or the number of factors is increased. This motivates us to develop a multi-objective criterion that minimizes the pairwise correlations as well as maximize the inter-site distances.

Because of the huge combinatorial nature of the problem, finding the optimal LHD is a very difficult task. Several algorithms such as simulated annealing (Morris and Mitchell 1995), columnwise-pairwise algorithms (Ye, Li, and Sudjianto 2000), enhanced stochastic evolutionary algorithms (Jin, Chen, and Sudjianto 2005), etc.

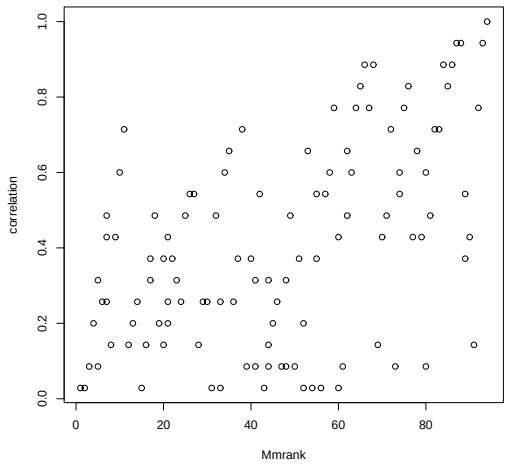

Figure 2: maximin rank vs correlation in $n = 6, k = 2$ case

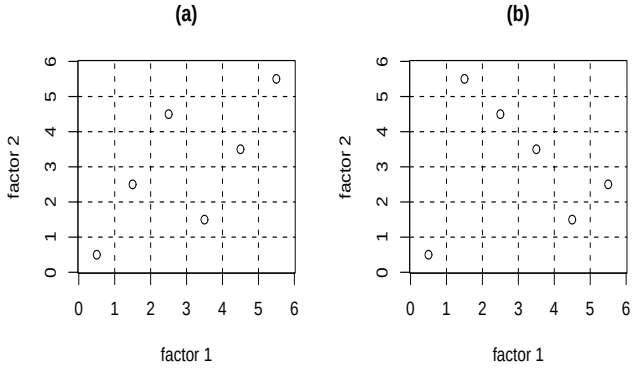

Figure 3: LHDs with $n = 6$ and $k = 2$. (a) correlation=0.714, maximin rank=11. (b) correlation=0.086, maximin rank=80.

are proposed in the literature for finding the optimal LHD. Most of the algorithms use an exchange method for searching in the design space. For example, in the algorithm proposed by Morris and Mitchell, a column in the design is randomly selected and then two randomly chosen elements within that column are exchanged to find a new design. We observed that the columns in the design matrix correspond to the experimental factors and thus we can choose them deterministically to reduce the pairwise correlations. Similarly, the rows in the design matrix correspond to the points in the experimental region and thus the elements can be chosen to maximize the inter-site distances. These observations lead to a new algorithm, which is highly suitable for finding the optimum based on our multi-objective criterion.

The chapter is organized as follows. In Section 2, performance measures for evaluating the goodness of an LHD with respect to pairwise correlations and inter-site distances are described. In Section 3, we propose a multi-objective criterion combining the two performance measures. In Section 4, we propose a new algorithm for generating optimal designs. Several examples are presented in Section 5 and a statistical justification for the new criterion is given in Section 6.

1.2 Performance Measures

Iman and Conover (1982), Owen (1994), and Tang (1998) proposed to choose designs by minimizing correlations among factors within the class of LHDs. We will use the following performance measure proposed by Owen, for evaluating the goodness of the LHD with respect to pairwise correlations. It is defined as

$$\rho^2 = \frac{\sum_{i=2}^{k}\sum_{j=1}^{i-1}\rho_{ij}^2}{k(k-1)/2}, \tag{1}$$

where ρ_{ij} is the linear correlation between columns i and j.

Now we will discuss a performance measure based on the inter-site distances. Let $\boldsymbol{X}$ be the design, which is an $n \times k$ matrix. Let $\boldsymbol{s}$ and $\boldsymbol{t}$ be any two design points (or sites). Consider the distance measure $d(\boldsymbol{s}, \boldsymbol{t}) = \{\sum_{j=1}^{k} |s_j - t_j|^p\}^{1/p}$, in which

$p = 1$ and $p = 2$ correspond to the rectangular and Euclidean distances respectively. Johnson, Moore, and Ylvisaker (1990) proposed the maximin distance criterion, which maximizes the minimum inter-site distance. Morris and Mitchell (1995) applied this criterion to the class of LHDs to find the optimal LHD. Because there are many designs that maximize the minimum inter-site distance, they proposed an extended definition of the maximin criterion. For a given LHD, define a distance list $(D_1, D_2, ..., D_m)$ in which the elements are the distinct values of inter-site distances, sorted from the smallest to the largest. Hence $m \leq \binom{n}{2}$. Let J_i be the number of pairs of sites in the design separated by D_i. Then a design $\boldsymbol{X}$ is called a maximin design if it sequentially maximizes D_i's and minimizes J_i's in the following order: $D_1, J_1, D_2, J_2, \cdots, D_m, J_m$. Morris and Mitchell (1995) then proposed a scalar-valued function which can be used to rank competing designs in such a way that the maximin design received the highest ranking. The family of functions indexed by p is given by

$$\phi_p = (\sum_{i=1}^{m} J_i D_i^{-p})^{1/p}, \tag{2}$$

where p is a positive integer. Now for large enough p, the design that minimizes ϕ_p will be a maximin design. In the next section we propose a new criterion which combines the performance measures in (5) and (2).

1.3 Multi-Objective Criterion

Our objective is to find an LHD that minimizes both ρ^2 and ϕ_p. A common approach in multi-objective optimization is to optimize a weighted average of all the objective functions. Therefore consider the objective function

$$w_1\rho^2 + w_2\phi_p,$$

where w_1 and w_2 are some pre-specified positive weights. Because the two objectives are very different, it is not easy to choose appropriate weights. Moreover, the two

objectives have different scales. The objective function $\rho^2 \in [0,1]$, whereas the objective function ϕ_p can be more than 1. If we can scale ϕ_p also to $[0,1]$, then we might be able to assign some reasonable weights. In order to do this, we need to find an upper and lower bound for ϕ_p. This is what we try to do in the following.

Consider an LHD with n points and k factors, denoted by $LHD(n,k)$. Suppose each factor takes values in $\{1,2,\cdots,n\}$. Let $d_1, d_2, \cdots, d_{\binom{n}{2}}$ be the inter-site distances among the n points based on the rectangular distance measure $d(\boldsymbol{s},\boldsymbol{t}) = \sum_{j=1}^{k} |s_j - t_j|$. We will use the following two results for deriving bounds for ϕ_p. All the proofs are given in the Appendix.

LEMMA 1. *For an $LHD(n,k)$, the average inter-site distance (rectangular measure) is a constant given by*

$$\overline{d} = \frac{(n+1)k}{3}.$$

LEMMA 2. *Consider a set of positive values $\{d_{j1}, d_{j2}, \cdots, d_{jm}\}$ and denote its ordered sequence by $d_{j(1)} \le d_{j(2)} \le \cdots \le d_{j(m)}$ for $j = 1,2,\cdots,k$. Then*

$$\sum_{i=1}^{m} \frac{1}{\sum_{j=1}^{k} d_{ji}} \le \sum_{i=1}^{m} \frac{1}{\sum_{j=1}^{k} d_{j(i)}}.$$

Lemma 1 shows that for all LHDs, the average distance is a constant for a given n and k. As an interesting consequence, note that the last step in the definition of maximin criterion cannot be applied to an LHD, because D_m is determined by $D_1, \cdots, D_{m-1}$. Therefore, it is more appropriate to define the objective function for the distances as $(\sum_{i=1}^{m-1} J_i D_i^{-p})^{1/p}$. But we will continue to use (2), because it has a computationally simpler form (Jin, Chen, Sudjianto, 2005). It can be written as

$$\phi_p = \left(\sum_{i=1}^{\binom{n}{2}} \frac{1}{d_i^p} \right)^{1/p},$$

which can be easily calculated (no need to order the d_i's) .

Let

$$\phi_{p,L} = \left\{ \binom{n}{2} \left(\frac{\lceil \bar{d} \rceil - \bar{d}}{\lfloor \bar{d} \rfloor^p} + \frac{\bar{d} - \lfloor \bar{d} \rfloor}{\lceil \bar{d} \rceil^p} \right) \right\}^{1/p} \quad \text{and} \quad \phi_{p,U} = \left\{ \sum_{i=1}^{n-1} \frac{(n-i)}{(ik)^p} \right\}^{1/p},$$

where $\lfloor x \rfloor$ is the largest integer $\leq x$ and $\lceil x \rceil$ is the smallest integer $> x$. The following result states that the above two values can be used as a lower and upper bound for ϕ_p.

PROPOSITION 1. *For an LHD(n, k), $\phi_{p,L} \leq \phi_p \leq \phi_{p,U}$.*

It is easy to see that the upper bound is achieved when all of the factors are equal. Thus the worst design in terms of ϕ_p is the same as the worst design in terms of ρ. However, there may not exist a design that achieves the lower bound.

Thus $(\phi_p - \phi_{p,L})/(\phi_{p,U} - \phi_{p,L}) \in [0, 1]$ has the same range as ρ^2. Therefore, our new criterion is to minimize

$$\psi_p = w\rho^2 + (1 - w)\frac{\phi_p - \phi_{p,L}}{\phi_{p,U} - \phi_{p,L}},$$

where $w \in (0, 1)$. The case of $w = 0.5$ gives approximately equal importance to both the correlation and the distance measures. We will call a design that minimizes ψ_p as an orthogonal-maximin Latin hypercube design (OMLHD). In the next section we propose a new algorithm for finding an OMLHD.

1.4 A New Algorithm

Morris and Mitchell (1995) proposed a version of the simulated annealing algorithm for optimizing ϕ_p. We will call their algorithm as MMA. In MMA, the search begins with a randomly chosen LHD, and proceeds through the examination of a sequence of designs, each generated as a perturbation of the preceding one. A perturbation $\boldsymbol{X}_{try}$ of a design $\boldsymbol{X}$ is generated by interchanging two randomly chosen elements within a randomly chosen column in $\boldsymbol{X}$. The perturbation $\boldsymbol{X}_{try}$ replaces $\boldsymbol{X}$ if it leads to an

improvement. Otherwise, it will replace $\boldsymbol{X}$ with probability $\pi = exp\{-[\phi_p(\boldsymbol{X}_{try}) - \phi_p(\boldsymbol{X})]/t\}$, where t is a preset parameter known as "temperature".

We propose a modification of the above algorithm. Instead of randomly choosing a column and two elements within that column, we will choose them judiciously in order to achieve improvement in our multi-objective function. Suppose at some stage of the iterations, a column is almost orthogonal to the other columns. Then clearly, we will not gain much in perturbing this column. It is much better to choose a column that is highly correlated with the other columns, because through a perturbation of its elements we may be able to reduce the correlation, thereby improving our objective function. Similarly, if a point is far from the other points, then there is no need to perturb the elements in that row. Instead, we can choose a point that is close to the other points and perturb the elements in the chosen column. This may increase the distance of the point from the others, thereby improving our objective function. For doing this, at each step, compute

$$\rho_l^2 = \frac{1}{k-1} \sum_{j \neq l} \rho_{lj}^2, \tag{3}$$

for each column $l = 1, 2, \cdots, k$ and

$$\phi_{pi} = \left(\sum_{j \neq i} 1/d_{ij}^p \right)^{1/p}, \tag{4}$$

for each row $i = 1, 2, \cdots, n$, where ρ_{lj} is the correlation between columns l and j; and d_{ij} is the distance between the rows i and j. Thus ρ_l^2 and ϕ_{pi} can be used as measures for correlation and distance for each column and row respectively. For exchanging the elements, we want to choose a column with high probability that is highly correlated with the other columns. Similarly, we want to choose a row with high probability that is closest to the other rows. Therefore choose the column

$$l^* = l \text{ with probability } P(l) = \frac{\rho_l^\alpha}{\sum_{l=1}^k \rho_l^\alpha},$$

and the row

$$i^* = i \text{ with probability } P(i) = \frac{\phi_{pi}^{\alpha}}{\sum_{i=1}^{n} \phi_{pi}^{\alpha}},$$

with $\alpha \in [1, \infty)$. Note that if ρ_l (or ϕ_{pi}) is high for a column (or row), then it will be chosen with a higher probability than the other columns (or rows). This step makes our algorithm different from the existing algorithms. Now exchange $x_{i^*l^*}$ with a randomly chosen element $x_{i'l^*}$. This gives us the new design $\boldsymbol{X}_{try}$. If $\psi_p(\boldsymbol{X}_{try}) < \psi_p(\boldsymbol{X})$, then we will replace $\boldsymbol{X}$ by $\boldsymbol{X}_{try}$, otherwise we will replace it with probability $\pi = exp\{-[\psi_p(\boldsymbol{X}_{try}) - \psi_p(\boldsymbol{X})]/t\}$.

All the parameters in the new algorithm are set the same as that used in a standard simulated annealing algorithm for which the convergence is already established (Lundy and Mees 1986). Therefore the new algorithm will also converge to the global optimum. A limiting case of the algorithm is interesting. When $\alpha \to \infty$, the exchange rule becomes deterministic, given by

$$l^* = arg\max_{l} \rho_l^2 \text{ and } i^* = arg\max_{i} \phi_{pi}.$$

Under this rule, the transition probability matrix for moving from one design to another design can be reducible, violating one of the conditions required for convergence. But our simulations, given in the next section, show that the convergence is faster with the above modification. Therefore, we recommend it for use in practice.

Because the objective function is evaluated at each iteration of the algorithm, it is extremely important to have a computationally efficient implementation of the objective function (see Jin, Chen, and Sudjianto 2005). Instead of calculating ρ_l^2 and ϕ_{pi} using (3) and (4), we can use the following iterative formulas. Let $(\rho_l^2)^{(s)}$ and $\phi_p^{(s)}$ denote the values of ρ_l^2 and ϕ_p at the iteration step s. Then at step $(s+1)$

$$\phi_{pi}^{(s+1)} = \begin{cases} \left(\sum_{j \neq i} 1/(d_{ij}^{(s+1)})^p\right)^{1/p} & , i = i', i^* \\ \left((\phi_{pi}^{(s)})^p - (d_{i^*i}^{(s)})^{-p} - (d_{i'i}^{(s)})^{-p} + (d_{i^*i}^{(s+1)})^{-p} + (d_{i'i}^{(s+1)})^{-p}\right)^{1/p} & , i \neq i', i^* \end{cases}.$$

For all $j \neq i^*, i'$ we have $d_{i^*j}^{(s+1)} = d_{i^*j}^{(s)} - t(i^*, i', j, l^*)$, and $d_{i'j}^{(s+1)} = d_{i'j}^{(s)} + t(i^*, i', j, l^*)$, where $t(i_1, i_2, u, v) = |x_{i_1 v} - x_{uv}| - |x_{i_2 v} - x_{uv}|$. Also note that the distance matrix (d_{ij}) is symmetric. For ρ_l^2 at step $(s+1)$, we obtain

$$(\rho_l^2)^{(s+1)} = \begin{cases} \frac{1}{k-1} \sum_{j \neq l} (\rho_{jl}^2)^{(s+1)} & , l = l^* \\ (\rho_l^2)^{(s)} + \frac{(\rho_{ll^*}^2)^{(s+1)} - (\rho_{ll^*}^2)^{(s)}}{k-1} & , l \neq l^* \end{cases}.$$

Thus

$$\phi_p^{(s+1)} = \left(\frac{1}{2} \sum_{i=1}^{n} (\phi_{pi}^{(s+1)})^p \right)^{1/p} \quad \text{and} \quad (\rho^2)^{(s+1)} = (\rho^2)^{(s)} + \frac{2(\rho_{l^*}^2)^{(s+1)} - 2(\rho_{l^*}^2)^{(s)}}{k}.$$

We should point out that the proposed exchange procedure can also be implemented with any of the other stochastic optimization algorithms such as the columnwise-pairwise algorithm (Li and Wu 1997, Ye, Li, and Sudjianto 2000), the threshold accepting heuristic (Winker and Fang 1998), and the stochastic evolutionary algorithm (Jin, Chen, and Sudjianto 2005).

1.5 Examples

In this section, we compare our proposed method with some of the existing methods. For a fair comparison, we choose all the parameters in the simulated annealing algorithm equal to the recommended values in Morris and Mitchell (1995). In the following examples, we let $p = 15$ and $w = 0.5$. In all the examples, we started the iteration by using a randomly generated symmetric LHD (Ye, Li, and Sudjianto 2000).

Example 1 (*OMLHD vs MLHD*). Consider an $LHD(5, 3)$. In this case it is feasible to enumerate all the LHDs. We found that there are a total of 142 different designs according to the maximin criterion (Morris and Mitchell, 1995). The maximin Latin hypercube design (MLHD) and the proposed OMLHD are given in Table 7. We see that for OMLHD, the maximum pairwise correlation is only 0.1 compared to 0.4 of MLHD. The minimum inter-site distances of the two designs are the same ($D_1 = 5$), although the number of sites separated by this distance is one less in MLHD.

Table 1: Example 1, MLHD vs OMLHD for $n = 5$ and $k = 3$

	MLHD	OMLHD
optimal design matrix	1 1 2 2 5 3 3 2 5 4 3 1 5 4 4	1 2 3 2 4 5 3 5 1 4 1 2 5 3 4
ϕ_p	0.2170	0.2201
$D_1(J_1)$	5(3)	5(4)
ρ	0.265	0.081
pairwise correlations	(0.4,0.2,0.1)	(-0.1,-0.1,0)

Table 2: Examples 2 and 4, OMLHD vs OLHD, MLHD, and ULHD for $n = 9$ and $k = 4$

	MLHD	OMLHD	OLHD	ULHD
optimal design matrix	1 3 3 4 2 5 8 8 3 8 6 2 4 7 1 6 5 2 9 3 6 9 5 9 7 1 4 7 8 4 2 1 9 6 7 5	1 5 3 3 2 2 5 8 3 9 7 5 4 3 8 1 5 7 1 7 6 6 9 9 7 1 2 4 8 8 4 2 9 4 6 6	1 2 6 3 2 9 7 6 3 4 2 9 4 7 1 2 5 5 5 5 6 3 9 8 7 6 8 1 8 1 3 4 9 8 4 7	4 1 7 5 1 3 4 3 9 9 5 4 6 6 6 9 5 7 2 1 2 8 8 7 3 5 1 6 8 2 3 8 7 4 9 2
ϕ_p	0.1049	0.1049	0.1154	0.1127
$D_1(J_1)$	11(3)	11(4)	10(8)	10(5)
ρ	0.108	0.063	0	0.076
maximum pairwise correlation	0.217	0.117	0	0.15
CL_2	0.1415	0.1386	0.1457	0.1374

Example 2 (*OMLHD vs OLHD*). Ye (1998) proposed the orthogonal Latin hypercube designs (OLHD), in which all the columns are orthogonal (correlation = 0) to each other. Table 2 compares the OLHD with the proposed OMLHD for the case of $n = 9$ and $k = 4$. For comparison, the MLHD is also given in the Table. We can see that the OMLHD is a compromise between the MLHD and OLHD. OLHD exists only for certain n and k, whereas MLHD and OMLHD exist for all n and k. In this sense MLHD and OMLHD are more general.

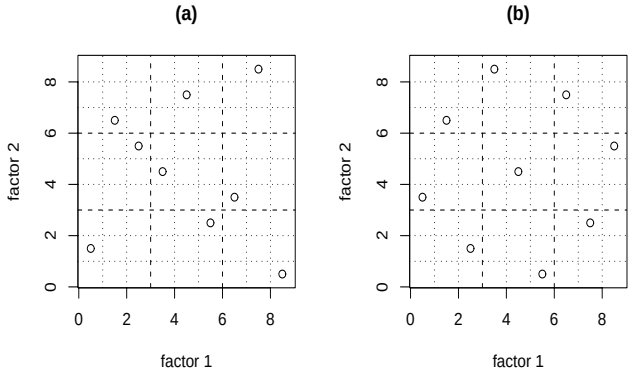

Figure 4: Example 3, (a) OA-based LHD ($\phi_p = 0.5380, D_1(J_1) = 2(3), \rho = -0.067$) (b)OMLHD ($\phi_p = 0.2879, D_1(J_1) = 4(8), \rho = 0$)

Example 3 (*OMLHD vs OA-based LHD*). Owen (1992) and Tang (1993) proposed using orthogonal arrays for constructing good LHDs. Tang called such designs OA-based LHDs. Figure 4 shows an OA-based LHD and the OMLHD for the case of $n = 9$ and $k = 2$. Clearly the OMLHD is superior to this particular OA-based LHD. Interestingly, in this case, the OMLHD is also an OA-based LHD, but a good one in terms of both correlation and space-filling. However, in general an OMLHD need not be an OA-based LHD.

Example 4 (*OMLHD vs ULHD*). Another popular space-filling design is the unform design. It can be obtained by minimizing the centered L_2-discrepancy criterion (CL_2)(see Fang, Ma, and Winker 2000). Denote the optimal LHD under this criterion by ULHD. The ULHD for $n = 9$ and $k = 4$ is given in Table 2. We can see that the

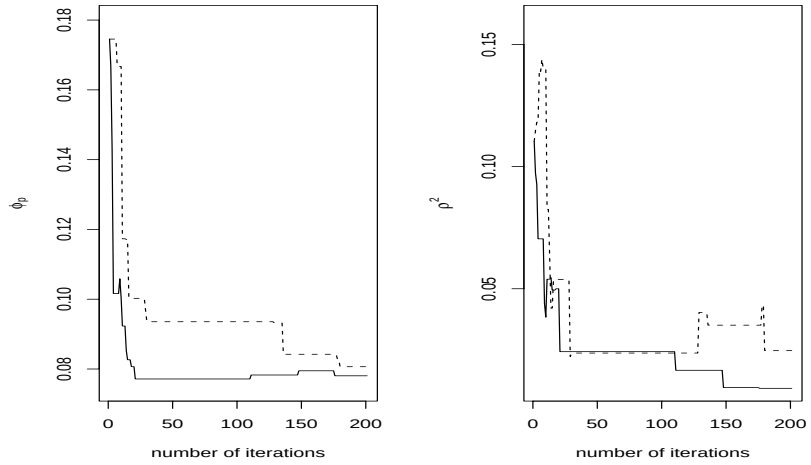

Figure 5: Performance of our new algorithm (solid) and MMA (dashed) against the number of iterations.

OMLHD is slightly worse than the ULHD under this criterion, but is better in terms of both ϕ_p and ρ. Interestingly, the OMLHD performs much better than MLHD and OLHD in terms of CL_2.

We have also studied the performance of the proposed exchange algorithm. Figure 5 shows how ϕ_p and ρ^2 are reduced with each iteration for the case of $LHD(25, 4)$. The same starting design is used for both MMA and the new algorithm. We can see that the new algorithm converges more quickly than the MMA. We repeated this 200 times. The values of ψ_p at the 50^{th} iteration are plotted in Figure 26. We can see that they are much smaller for the new algorithm compared to the MMA. Thus for a fixed number of iterations, the new algorithm produces LHDs with smaller pairwise correlations and larger inter-site distances. The simulations are repeated for $LHD(50, 4)$, $LHD(10, 10)$, and $LHD(100, 10)$. The number of iterations for each of these cases was fixed at 100, 200, and 500 respectively. The results are shown in Figure 26. We can see that remarkable improvements are obtained by using the new algorithm compared to the MMA.

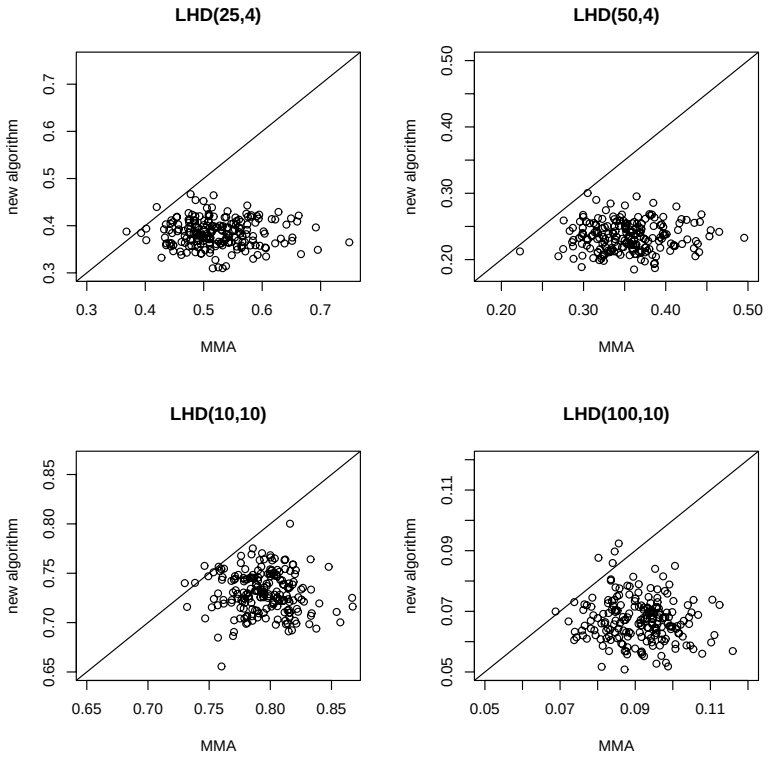

Figure 6: Plot of ψ_p values from the new algorithm against that of the MMA.

1.6 A Statistical Justification

Because of the absence of random errors, interpolating methods such as kriging are widely used for modeling and analysis in computer experiments. Consider a function $y(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, \cdots, x_k)'$. The ordinary kriging model is given by,

$$Y(\boldsymbol{x}) = \mu + Z(\boldsymbol{x}), \tag{5}$$

where $Z(\boldsymbol{x})$ is a weak stationary stochastic process with mean 0 and covariance function $\sigma^2 R$. A popular choice for the correlation function is the exponential correlation function:

$$R(\boldsymbol{h}) = e^{-\theta \sum_{i=1}^{k} |h_i|^{\gamma}}, \tag{6}$$

with $\theta \in (0, \infty)$ and $\gamma \in (0, 2]$. Suppose we evaluated the function at n points $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ and let $\boldsymbol{y} = (y_1, \cdots, y_n)'$ be the corresponding function values. Then, the best linear unbiased predictor (BLUP) is given by $\hat{y}(\boldsymbol{x}) = \hat{\mu} + \boldsymbol{r}(\boldsymbol{x})' \boldsymbol{R}^{-1} (\boldsymbol{y} - \hat{\mu}\boldsymbol{1})$,

where $\mathbf{1}$ is a column of 1's having length n, $\boldsymbol{r}(\boldsymbol{x})' = (R(\boldsymbol{x}-\boldsymbol{x}_1), \cdots, R(\boldsymbol{x}-\boldsymbol{x}_n))$, $\boldsymbol{R}$ is an $n \times n$ matrix with elements $R(\boldsymbol{x}_i - \boldsymbol{x}_j)$, and $\hat{\mu} = \mathbf{1}'\boldsymbol{R}^{-1}\boldsymbol{y}/\mathbf{1}'\boldsymbol{R}^{-1}\mathbf{1}$. Note that the model in (2) assumes a constant mean and therefore, the predictor does not perform well when there are some trends (see Joseph (2006a)). If the trends are known, then universal kriging can be used instead of ordinary kriging. The universal kriging model with linear trends is given by

$$Y(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + Z(\boldsymbol{x}), \tag{7}$$

where $\beta_0, \beta_1, \cdots, \beta_k$ are some unknown constants. Simulations carried out by Martin and Simpson (2005) show that universal kriging can improve the prediction over ordinary kriging. See Qian et al. (2006) for a real application of universal kriging with linear trends.

Johnson, Moore, and Ylvisaker (1990) have shown that the maximin design with minimum J_1 is asymptotically D-optimum under the ordinary kriging model (as correlation becomes weak). Thus the objective of a maximin design can be thought of as finding a design to improve prediction through the stochastic part $Z(\boldsymbol{x})$. Whereas minimizing the correlation among the variables will help in estimating the deterministic mean part $\beta_0 + \sum_{i=1}^{k} \beta_i x_i$ efficiently. For the universal kriging predictor to perform well, both parts need to be estimated precisely. Thus the orthogonal-maximin LHD can be considered suitable for the efficient estimation of the universal kriging model with linear trends.

More specifically, consider the following hierarchical Bayesian model:

$$\boldsymbol{y}|\boldsymbol{\beta} \sim N(\boldsymbol{F}\boldsymbol{\beta}, \sigma^2\boldsymbol{R}), \quad \boldsymbol{\beta} \sim N(\boldsymbol{\mu}, \tau^2\boldsymbol{I}),$$

where $\boldsymbol{F} = [\mathbf{1}, \boldsymbol{X}]$ is the model matrix corresponding to $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_k)'$ and $\boldsymbol{I}$ is an identity matrix. The maximum entropy design (Shewry and Wynn 1987) is obtained by maximizing the determinant of the variance-covariance matrix of $\boldsymbol{y}$. Thus we need to maximize $det(\sigma^2\boldsymbol{R} + \tau^2\boldsymbol{F}\boldsymbol{F}')$, which is equal to (see Santner, Williams, and Notz

2003, page 167)

$$det(\sigma^2 \boldsymbol{R})\ det(\tau^2/\sigma^2 \boldsymbol{F}' \boldsymbol{R}^{-1} \boldsymbol{F} + \boldsymbol{I}).$$

Johnson, Moore, and Ylvisaker (1990) have shown that as $\theta \to \infty$ in (6), a maximin design maximizes the first term $det(\sigma^2 \boldsymbol{R})$. As $\theta \to \infty$, $\tau^2/\sigma^2 \boldsymbol{F}' \boldsymbol{R}^{-1} \boldsymbol{F} + \boldsymbol{I} \to \tau^2/\sigma^2 \boldsymbol{F}' \boldsymbol{F} + \boldsymbol{I}$, whose determinant is maximized when $\boldsymbol{F}$ is orthogonal. Thus an orthogonal design maximizes the second term. A design will be asymptotically ($\theta \to \infty$) optimum with respect to the maximum entropy criterion if both the terms are maximized. Therefore, an OMLHD, which possesses good maximin and orthogonality properties, can be expected to perform well in terms of the maximum entropy criterion for the model in (1) among all LHDs.

CHAPTER II

BLIND KRIGING: A NEW METHOD FOR DEVELOPING METAMODELS

[1]

2.1 Introduction

The use of computer modeling and experiments is becoming more and more popular for product design optimization (Fang, Li, and Sudjianto, 2006). Based on the physical knowledge of the product, models such as finite element models can be formulated and solved on computers. Although cheaper than experimenting on products or prototypes, computer experiments can still be time consuming and expensive. An approach to reduce the computational time and cost is to perform optimization on a metamodel that approximates the original computer model. The metamodel can be estimated from data by running the computer experiment on a sample of points in the region of interest.

Kriging is widely used for obtaining the metamodels (Sacks, et al., 1989; Santner, Williams, and Notz, 2003; Jin, Chen, and Simpson, 2001). For examples, Pacheco, Amon, and Finger (2003) uses kriging for the thermal design of wearable computers, Cappelleri, et al. (2002) uses kriging for the design of a variable thickness piezoelectric bimorph actuator, and so on. The popularity of kriging is due to the fact that computer models are often deterministic (i.e., no random error in the output) and thus interpolating metamodels are desirable. Kriging gives an interpolating metamodel and is therefore more suitable than the other common alternatives such as quadratic

[1]The paper based on this chapter appears in *ASME Journal of Mechanical Design*, **130**, 031102, 2008.

response surface model.

A kriging model, known as *universal kriging*, can be stated as follows (Wackernagel, 2002). Assume that the true function $y(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^p$, is a realization from a stochastic process

$$Y(\boldsymbol{x}) = \mu(\boldsymbol{x}) + Z(\boldsymbol{x}), \tag{1}$$

where $\mu(\boldsymbol{x}) = \sum_{i=0}^{m} \mu_i v_i(\boldsymbol{x})$ and $Z(\boldsymbol{x})$ is a weak stationary stochastic process with mean 0 and covariance function $\sigma^2\psi$. The v_i's are some known functions and μ_i's are unknown parameters. Usually $v_0(\boldsymbol{x}) = 1$. The covariance function is defined as $cov\{Y(\boldsymbol{x}+\boldsymbol{h}), Y(\boldsymbol{x})\} = \sigma^2\psi(\boldsymbol{h})$, where the correlation function $\psi(\boldsymbol{h})$ is a positive semidefinite function with $\psi(\boldsymbol{0}) = 1$ and $\psi(-\boldsymbol{h}) = \psi(\boldsymbol{h})$. In this formulation $\mu(\boldsymbol{x})$ is used to capture the known trends, so that $Z(\boldsymbol{x})$ will be a stationary process. But, in reality, rarely will those trends be known and thus the following special case, known as *ordinary kriging*, is commonly used (Wackernagel, 2002; Currin, et al., 1991; Welch, et al., 1992),

$$Y(\boldsymbol{x}) = \mu_0 + Z(\boldsymbol{x}). \tag{2}$$

The metamodel (or the predictor) can be obtained as follows. Suppose we evaluated the function at n points $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ and let $\boldsymbol{y} = (y_1, \cdots, y_n)'$ be the corresponding function values. Then ordinary kriging predictor is given by

$$\hat{y}(\boldsymbol{x}) = \hat{\mu_0} + \boldsymbol{\psi}(\boldsymbol{x})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \hat{\mu_0}\mathbf{1}), \tag{3}$$

where $\mathbf{1}$ is a column of 1's having length n, $\boldsymbol{\psi}(\boldsymbol{x})' = (\psi(\boldsymbol{x}-\boldsymbol{x}_1), \cdots, \psi(\boldsymbol{x}-\boldsymbol{x}_n))$, $\boldsymbol{\Psi}$ is an $n \times n$ matrix with elements $\psi(\boldsymbol{x}_i - \boldsymbol{x}_j)$, and $\hat{\mu_0} = \mathbf{1}'\boldsymbol{\Psi}^{-1}\boldsymbol{y}/\mathbf{1}'\boldsymbol{\Psi}^{-1}\mathbf{1}$. It is the best linear unbiased predictor, which minimizes the mean squared prediction error $E\{\hat{Y}(\boldsymbol{x}) - Y(\boldsymbol{x})\}^2$ under the model in Eq. (2).

The predictor in Eq. (3) is an interpolating predictor and is easy to evaluate. However, it has some problems. First, the prediction can be poor if there are some strong trends (see the simulation results in Martin and Simpson (2005). Second, it

is not easy to understand the effects of the factors by just looking at the predictor. Of course, sensitivity analysis techniques such as the functional analysis of variance can be used for understanding and quantifying their effects (Fang, Li, and Sudjianto, 2006), but the proposed predictor in this chapter is a much simpler alternative. Third, the predictor is not robust to the misspecification in the correlation parameters (see Joseph (2006a) for examples). In this chapter, we propose a modification of universal kriging predictor that overcomes the foregoing problems observed with ordinary kriging predictor.

2.2 Blind Kriging

We propose a simple modification to universal kriging model in Eq. (1). We do not assume the functions v_i's to be known. Instead, they are identified through some data-analytic procedures. Because v_i's are unknown in our model, we name it *blind kriging*. Thus the blind kriging model is given by

$$Y(\boldsymbol{x}) = \boldsymbol{v}(\boldsymbol{x})'\boldsymbol{\mu}_m + Z(\boldsymbol{x}), \tag{4}$$

where $\boldsymbol{v}(\boldsymbol{x})' = (1, v_1, \cdots, v_m)$, $\boldsymbol{\mu}_m = (\mu_0, \mu_1, \cdots, \mu_m)'$, and m are unknown. Here $Z(\boldsymbol{x})$ is assumed to be a weak stationary stochastic process with mean 0 and covariance function $\sigma_m^2\psi$. The correlation function ψ can also depend on m, but for the moment assume it to be independent. The blind kriging predictor, which has the same form as that of universal kriging predictor, is given by

$$\hat{y}(\boldsymbol{x}) = \boldsymbol{v}(\boldsymbol{x})'\hat{\boldsymbol{\mu}}_m + \boldsymbol{\psi}(\boldsymbol{x})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{V}_m\hat{\boldsymbol{\mu}}_m), \tag{5}$$

where $\boldsymbol{V}_m = (\boldsymbol{v}(\boldsymbol{x}_1), \cdots, \boldsymbol{v}(\boldsymbol{x}_n))'$ and $\hat{\boldsymbol{\mu}}_m = (\boldsymbol{V}_m'\boldsymbol{\Psi}^{-1}\boldsymbol{V}_m)^{-1}\boldsymbol{V}_m'\boldsymbol{\Psi}^{-1}\boldsymbol{y}$. Note that $\boldsymbol{V}_m$ is an $n \times (m+1)$ matrix.

The most important step in blind kriging is to identify the unknown functions v_i's. They can be chosen from a set of candidate functions (or variables) using variable selection techniques. If some simple functions are used in the candidate set, then the

predictor can be easily interpreted using the first part $\boldsymbol{v}(\boldsymbol{x})'\hat{\boldsymbol{\mu}}_m$. The second part of the predictor $\boldsymbol{\psi}(\boldsymbol{x})'\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{V}_m\hat{\boldsymbol{\mu}}_m)$ is used to achieve interpolation and does not provide much information about the overall trend in the function.

2.2.1 Variable Selection

There are many variable selection techniques that are popular in regression analysis such as forward selection, backward elimination, and step-wise regression (Miller, 2002). Recently, many other techniques have also been proposed (George and McCulloch, 1993; Breiman, 1995; Tibshirani, 1996; Efron, et al. 2004). All of these techniques have a drawback for using in the analysis of experiments and in particular for blind kriging. They do not lead to models that satisfy the well known principles of effect hierarchy and effect heredity (Hamada and Wu, 1992). The *effect hierarchy principle* states that lower order effects (such as main effects) are more important than higher order effects (such as two-factor interactions) and the *effect heredity principle* states that in order for an interaction effect to be significant, at least one of its parent factors should be significant. These principles are useful for identifying models that are simple and interpretable. Chipman, Hamada, and Wu (1997) introduced a Bayesian variable selection technique that incorporates these two principles. Another Bayesian variable selection technique introduced in Joseph (2006b) and Joseph and Delaney (2007) seems to be more useful for our purpose because of its connections with kriging. It can be considered as a Bayesian version of the forward selection strategy. Below we explain this technique briefly. Additional details of the technique are included in the Appendix. We note that the work in Joseph (2006b) and Joseph and Delaney (2007) focus on physical experiments and therefore, the Bayesian variable selection technique was applied only to linear models and not kriging models.

The candidate variables are selected as the linear effects, quadratic effects, and two-factor interactions. Here the two-factor interactions include the linear-by-linear,

linear-by-quadratic, quadratic-by-linear, and quadratic-by-quadratic interactions. There are a total of $t = 2p^2$ candidate variables (excluding the constant term). We note that this Bayesian variable selection technique can easily handle three and higher order effects, but in this chapter we focus on the lower order effects for the simplicity of exposition and interpretation. Following Joseph and Delaney (2007), first scale the factors in $[1.0, 3.0]$. Other ranges such as $[0, 1]$ or $[-1, 1]$ maybe used, however, Eqs (6) and (7) should be changed accordingly (see the Appendix). The linear and quadratic effects can be defined using the orthogonal polynomial coding (Wu and Hamada, 2000)

$$x_{jl} = \frac{\sqrt{3}}{\sqrt{2}}(x_j - 2) \text{ and } x_{jq} = \frac{1}{\sqrt{2}}(3(x_j - 2)^2 - 2), \tag{6}$$

for $j = 1, 2, \cdots, p$. The variables x_{jl} and x_{jq} are scaled so that they have the same length $\sqrt{3}$ when x_j takes the values 1, 2, and 3. The two-factor interaction terms can be defined as the products of these variables. For example, the linear-by-quadratic interaction term between x_1 and x_3 can be defined as $x_{1l}x_{3q}$.

Denote the candidate variables by $u_1, \cdots, u_t$. Consider approximating $y(\boldsymbol{x})$ by the linear model $\sum_{i=0}^{m} \mu_i v_i + \sum_{i=0}^{t} \beta_i u_i$, where $u_0 = 1$. As an example, for two factors x_1 and x_2, the linear model is $\sum_{i=0}^{m} \mu_i v_i + \sum_{i=0}^{8} \beta_i u_i$, where $u_0 = 1$, $u_1 = x_{1l}$, $u_2 = x_{1q}$, $u_3 = x_{2l}$, $u_4 = x_{2q}$, $u_5 = x_{1l}x_{2l}$, $u_6 = x_{1l}x_{2q}$, $u_7 = x_{1q}x_{2l}$, and $u_8 = x_{1q}x_{2q}$. Note that when $t > n - 1$, a frequentist estimation of the β_i's is not possible. However, all of the t effects can be simultaneously estimated using a Bayesian approach. For doing this, we need to postulate a prior distribution for $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_t)'$. Let

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \tau_m^2 \boldsymbol{R}),$$

where $\mathbf{0}$ is a vector of 0's having length $t + 1$ and $\boldsymbol{R}$ is a $(t + 1) \times (t + 1)$ diagonal matrix.

The matrix $\boldsymbol{R}$ can be constructed as follows. Assume that the correlation function in ordinary kriging model has a product correlation structure given by $\psi(\boldsymbol{h}) =$

$\prod_{j=1}^{p} \psi_j(h_j)$. Let $l_{ij} = 1$ if β_i includes the linear effect of factor j and 0 otherwise. Similarly, $q_{ij} = 1$ if β_i includes the quadratic effect of factor j and 0 otherwise. Then the ith diagonal element of $\boldsymbol{R}$ is given by $\prod_{j=1}^{p} r_{jl}^{l_{ij}} r_{jq}^{q_{ij}}$, where

$$r_{jl} = \frac{3 - 3\psi_j(2)}{3 + 4\psi_j(1) + 2\psi_j(2)} \text{ and } r_{jq} = \frac{3 - 4\psi_j(1) + \psi_j(2)}{3 + 4\psi_j(1) + 2\psi_j(2)}. \tag{7}$$

The foregoing connection with kriging makes this Bayesian variable selection technique the most suitable among its competitors. As shown in Joseph (2006b) and Joseph and Delaney (2007), the effect hierarchy and effect heredity principles are embedded in the prior.

Assume that $Z(\boldsymbol{x})$ in Eq. (4) follows a Gaussian process. Then the posterior mean of $\boldsymbol{\beta}$ can be approximated by Joseph and Delaney (2007)

$$\hat{\boldsymbol{\beta}} = \frac{\tau_m^2}{\sigma_m^2} \boldsymbol{R}\boldsymbol{U}'\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{V}_m\hat{\boldsymbol{\mu}}_m), \tag{8}$$

where $\boldsymbol{U}$ is the model matrix corresponding to the experimental design. A variable can be declared important if its absolute coefficient is large. Thus the variable to enter at each step $m = 0, 1, 2, \cdots$ can be selected as the variable with the largest $|\hat{\beta}_i|$. We note that Joseph (2006b) and Joseph and Delaney (2007) instead uses the standardized coefficient for variable selection. Both produce similar results, but the computation of the former is easier. For maximizing $|\hat{\beta}_i|$, without loss of generality we can set $\tau_m^2/\sigma_m^2 = 1$ in Eq. (8), which significantly simplifies the computations.

There remains an important issue to address in this Bayesian forward variable selection strategy. When should we stop adding terms to the mean part? In other words, what is the best value for m? The difficulty in choosing m is that, irrespective of its value, kriging predictor interpolates the data and thus gives a perfect fit. Therefore, the prediction errors are all 0. This prevents us from using the standard model selection criteria in regression analysis such as C_p-statistic and Akaike information criterion (Miller 2002). We will overcome this problem by using cross validation errors.

Let $\hat{y}_{(i)}(\boldsymbol{x})$ be the predictor after removing the ith data point. Then the leave-one-out cross validation error is defined as

$$cv_i = y_i - \hat{y}_{(i)}(\boldsymbol{x}_i),$$

for $i = 1, 2, \cdots, n$. Define the cross validation prediction error (CVPE) by

$$CVPE(m) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} cv_i^2}.$$

Now we can choose the value of m that minimizes $CVPE(m)$. We should point out that the foregoing approach of using cross validation errors works well only if the experimental data points are able to capture the trends in the true function.

The cross validation errors can be computed only after estimating the unknown parameters from the data, which is discussed in the next section. Among the parameters, those associated with the correlation function are computationally difficult to estimate. Clearly, the computations will become even more difficult if we need to estimate those parameters after removing each data point. Therefore, we recommend keeping the correlation parameters the same when computing cross validation errors.

2.2.2 Estimation

We choose the following Gaussian product correlation function

$$\psi(\boldsymbol{h}) = \exp(-\sum_{j=1}^{p} \theta_j h_j^2),$$

which is the most popular correlation function used in computer experiments (Santner, Williams, and Notz, 2003). Let $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)'$. The parameters $\boldsymbol{\mu}_m$, σ_m^2, and $\boldsymbol{\theta}$ can be estimated by maximizing the likelihood. Because the model is selected based on a cross validation criterion, it may seem more appropriate to use the same criterion for estimation. However, many empirical studies have shown that the maximum likelihood estimates perform better than the estimates based on cross validation (Santner, Williams, and Notz, 2003; Martin and Simpson, 2005).

Under the assumption that $Z(\boldsymbol{x})$ in Eq. (4) follows a Gaussian process, the negative of the log-likelihood is given by

$$NL = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\sigma_m^2 + \frac{1}{2}\log|\boldsymbol{\Psi}| + \frac{1}{2\sigma_m^2}(\boldsymbol{y} - \boldsymbol{V}_m\boldsymbol{\mu}_m)'\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{V}_m\boldsymbol{\mu}_m).$$

For the moment assume that $\boldsymbol{\theta}$ is known. Minimizing NL with respect to $\boldsymbol{\mu}_m$ and σ_m^2, we obtain (Santner, Williams, and Notz, 2003)

$$\hat{\boldsymbol{\mu}}_m = (\boldsymbol{V}'_m\boldsymbol{\Psi}^{-1}\boldsymbol{V}_m)^{-1}\boldsymbol{V}'_m\boldsymbol{\Psi}^{-1}\boldsymbol{y}, \tag{9}$$

$$\hat{\sigma}_m^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{V}_m\hat{\boldsymbol{\mu}}_m)'\boldsymbol{\Psi}^{-1}(\boldsymbol{y} - \boldsymbol{V}_m\hat{\boldsymbol{\mu}}_m). \tag{10}$$

Thus the minimum value of NL is

$$NL = \frac{n}{2}(1 + \log(2\pi)) + \frac{1}{2}(n\log\hat{\sigma}_m^2 + \log|\boldsymbol{\Psi}|). \tag{11}$$

Now consider the case with unknown $\boldsymbol{\theta}$. It can also be estimated by minimizing NL in Eq. (11). However, the minimization is not a trivial task. We have encountered multiple local minima in many examples and thus, finding the global minimum is difficult. Therefore, we propose to estimate $\boldsymbol{\theta}$ only at $m = 0$. Thus

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \; n\log\hat{\sigma}_0^2 + \log|\boldsymbol{\Psi}|. \tag{12}$$

Keeping the correlation parameters the same at each step also helps in identifying a mean model that satisfies effect heredity Joseph and Delaney 2007. At the final step, that is after choosing m, the correlation parameters can be again estimated (i.e., by minimizing NL in Eq. (11)), which can give a better prediction. Because $\boldsymbol{\theta}$ is estimated two times, the computational complexity in fitting a blind kriging model is roughly twice as that of an ordinary kriging model. The approach is explained with examples in the next section.

2.3 Examples

2.3.1 Example 1: Engine Block and Head Joint Sealing Experiment

The engine block and head joint sealing assembly is one of the most crucial and fundamental structural design in the automotive internal combustion engine. Design

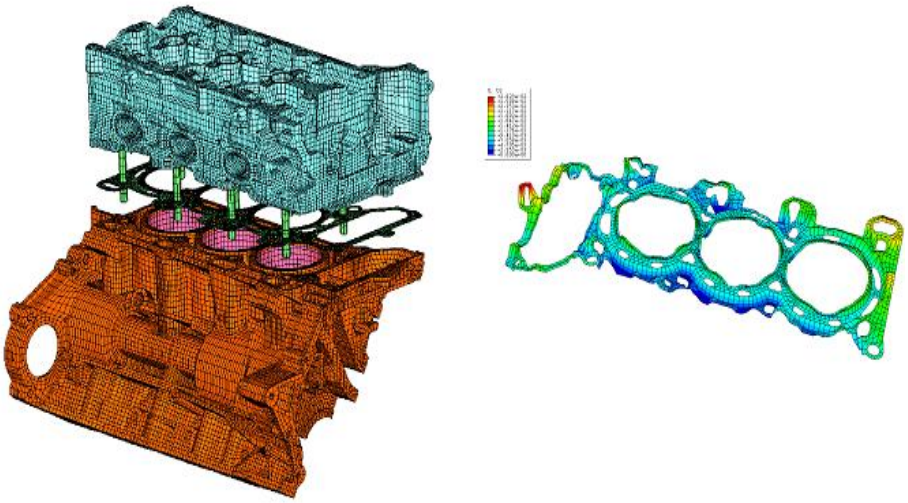

Figure 7: Finite element model of engine head and block joint sealing assembly

decisions must be made upfront, prior to the availability of a physical prototype, because it affects downstream design decisions for other engine components as well as significantly impacts the long lead time tooling and machining facility setup. Reversing a decision about this assembly at a later time has very expensive consequences. Thus, the use of a computer simulation model is indispensable. The design of the engine block and head joint sealing assembly is very complex due to multiple functional requirements (e.g., combustion gas, high pressure oil, oil drain, and coolant sealing) and complicated geometry; thus, the interactions among design parameters in this assembly (block and head structures, gasket, and fasteners) have significant effects. To best simulate the engine assembly process and operating conditions, a finite element model was developed to capture the complexity of part geometry, the compliance in the components, non-linear material properties, and contact interface between the parts (see Fig. 1). To address performance robustness of the joint sealing, manufacturing variability of the mating surfaces and head bolt tensional load are included in the analysis for which design parameters are optimized. Because the assembly model is computationally expensive, the availability of a computationally efficient and accurate metamodel is important for optimizing the design.

Table 3: Example 1, Data for the engine head and block joint sealing experiment

Run	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	y
1	2	2	3	2	2	1	2	3	1.53
2	3	3	3	2	3	1	3	1	2.21
3	1	1	2	3	2	1	3	3	1.69
4	3	1	2	1	2	2	3	1	1.92
5	1	1	2	2	3	1	1	2	1.42
6	1	3	2	3	3	3	2	2	5.33
7	1	3	1	2	1	2	3	3	2.00
8	2	3	2	1	1	1	1	1	2.13
9	3	2	1	3	3	2	1	2	1.77
10	2	1	1	2	1	3	1	3	1.89
11	1	3	3	1	3	2	1	3	2.17
12	3	2	2	3	1	2	1	3	2.00
13	3	3	1	3	2	1	2	3	1.66
14	2	1	1	3	3	2	3	1	2.54
15	1	2	1	1	3	1	2	1	1.64
16	3	1	3	2	3	3	2	3	2.14
17	1	2	3	1	1	3	3	2	4.20
18	3	2	2	2	1	3	2	1	1.69
19	1	2	1	2	2	3	1	1	3.74
20	2	2	2	1	3	3	3	3	2.07
21	2	3	3	3	2	3	1	1	1.87
22	2	3	2	2	2	2	2	2	1.19
23	3	3	1	1	2	3	3	2	1.70
24	2	2	3	3	1	1	3	2	1.29
25	2	1	1	1	1	1	2	2	1.82
26	1	1	3	3	1	2	2	1	3.43
27	3	1	3	1	2	2	1	2	1.91

Eight factors are selected for experimentation: gasket thickness (x_1), number of contour zones (x_2), zone-to-zone transition (x_3), bead profile (x_4), coining depth (x_5), deck face surface flatness (x_6), load/deflection variation (x_7), and head bolt force variation (x_8). Because of the complexity in the simulation setup and the excessive computing requirements, only 27 runs are used for the experiment. The experimental design, which is a 27-run orthogonal array (Wu and Hamada, 2000), is given in Table 3. In this example, we analyze only the gap lift (y).

First consider ordinary kriging. The maximum likelihood estimate of $\boldsymbol{\theta}$ is given by

$$\hat{\boldsymbol{\theta}} = (2.75, .26, .02, .01, .01, 4.00, .01, .01)'.$$

To avoid numerical problems, we have constrained each θ_i in $[.01, 4]$ in the optimization of the likelihood. We obtain $CVPE(0) = .5784$. The ordinary kriging predictor is given by

$$\hat{y}(\boldsymbol{x}) = 2.27 + \hat{\boldsymbol{\psi}}(\boldsymbol{x})'\hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{y} - 2.27\ \mathbf{1}),$$

where $\hat{\boldsymbol{\psi}}(\boldsymbol{x})$ is a vector of length 27 with ith element $\psi(\boldsymbol{x}-\boldsymbol{x}_i) = \exp(-\sum_{k=1}^{8}\hat{\theta}_k(x_k - x_{ik})^2)$ and $\hat{\boldsymbol{\Psi}}$ is a 27×27 matrix whose ijth element is $\psi(\boldsymbol{x}_i-\boldsymbol{x}_j) = \exp(-\sum_{k=1}^{8}\hat{\theta}_k(x_{ik} - x_{jk})^2)$.

Now consider blind kriging. To apply the Bayesian variable selection technique in Joseph (2006b) and Joseph and Delaney (2007), we first need to construct the $\boldsymbol{R}$ matrix. It is a 129×129 diagonal matrix given by

$$\hat{\boldsymbol{R}} = diag(1, \hat{r}_{1l}, \hat{r}_{1q}, \hat{r}_{2l}, \cdots, \hat{r}_{7q}\hat{r}_{8q}),$$

where

$$\hat{r}_{jl} = \frac{3 - 3e^{-4\hat{\theta}_j}}{3 + 4e^{-\hat{\theta}_j} + 2e^{-4\hat{\theta}_j}} \text{ and } \hat{r}_{jq} = \frac{3 - 4e^{-\hat{\theta}_j} + e^{-4\hat{\theta}_j}}{3 + 4e^{-\hat{\theta}_j} + 2e^{-4\hat{\theta}_j}}.$$

Now compute

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{R}}\boldsymbol{U}'\hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{y} - 2.27\ \mathbf{1}),$$

where $\boldsymbol{U}$ is a 27×129 matrix whose first column is $\mathbf{1}$ and the other columns correspond to the values of x_{1l}, x_{1q}, x_{2l}, $\cdots$, $x_{7q}x_{8q}$. Note that because we are only interested in finding the maximum value of $|\hat{\beta}_i|$, we have set $\tau_0^2/\sigma_0^2 = 1$ in Eq. (8). A half-normal plot (Wu and Hamada, 2000) of the absolute values of $\hat{\beta}_i$'s is shown in Fig. 8. We can see that the maximum value of $|\hat{\beta}_i|$ occurs for the coefficient of the linear-by-linear interaction term of x_1 and x_6. Note that to identify the largest absolute value of the coefficient, we do not need a half-normal plot; it is given here only for illustration.

Thus, take $v_1 = x_{1l}x_{6l}$. Again estimate the coefficients using

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{R}}\boldsymbol{U}'\hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{y} - \boldsymbol{V}_1\hat{\boldsymbol{\mu}}_1),$$

where $\hat{\boldsymbol{\mu}}_1$ is obtained from Eq. (9) and $\boldsymbol{V}_1$ is a 27×2 matrix whose first column is $\mathbf{1}$ and the second column is the values of v_1. Note that in this computation, the matrices $\hat{\boldsymbol{R}}$, $\boldsymbol{U}$, and $\hat{\boldsymbol{\Psi}}$ remain the same as before. At this step, we identify x_{1l} as the most significant among the remaining variables, because it has the largest $|\hat{\beta}_i|$. Thus, take $v_2 = x_{1l}$ and continue the forward selection procedure. In the next four steps, the variables x_{6l}, $x_{1q}x_{6l}$, x_{1q}, and $x_{2l}x_{6q}$ are selected. The $CVPE(m)$ decrease as shown in Fig. 9 (in the figure ordinary kriging is denoted by OK). The next variable to enter is x_{6q}, but it increases the $CVPE(m)$. We checked a few more steps and found that $CVPE(m)$ is continued to increase and thus we choose $m = 6$. We obtain $CVPE(6) = .4243$. It is also informative to calculate the usual R^2 value used in regression analysis. For our problem, we can define it by Joseph (2006b)

$$R^2(m) = 1 - \frac{\sum_{j=1}^{n}(y_j - \sum_{i=0}^{m}\hat{\mu}_i v_{ij})^2}{\sum_{j=1}^{n}(y_j - \hat{\mu}_0)^2}.$$

It is also plotted in Fig. 9. We can see that the six variables in the mean part explains about 86% of the variation in the data. The kriging part captures the remaining 14%.

The correlation parameters $\boldsymbol{\theta}$ can again be estimated by minimizing NL in Eq. (11). The new $\hat{\boldsymbol{\theta}}$ is obtained as

$$\hat{\boldsymbol{\theta}} = (.01, .01, .01, .01, 4, .24, 4, .14)'.$$

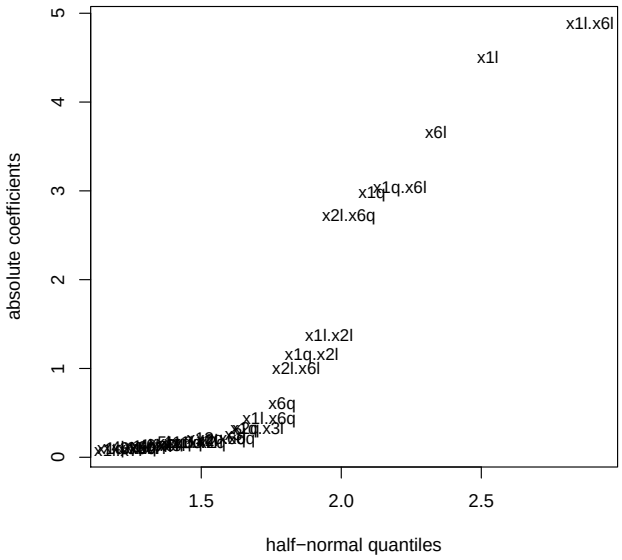

Figure 8: Half-normal plot of $|\hat{\beta}_i|$'s at $m = 0$

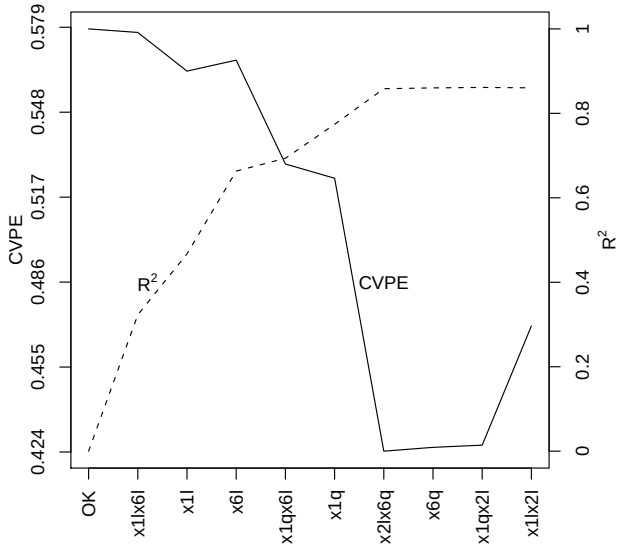

Figure 9: Plots of $CVPE(m)$ and $R^2(m)$ in Example 1

We obtain $CVPE(6) = .2702$, which is much smaller than using the $\boldsymbol{\theta}$ estimated at the beginning. The CVPE shows about 53% improvement in prediction using blind kriging over ordinary kriging ($CVPE(0) = .5784$).

The blind kriging predictor is given by

$$\hat{y}(\boldsymbol{x}) = 2.18 - .44x_{1l}x_{6l} - .48x_{1l} + .39x_{6l} + .21x_{1q}x_{6l} + .19x_{1q} + .30x_{2l}x_{6q} + \hat{\boldsymbol{\psi}}(\boldsymbol{x})'\hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{y} - \boldsymbol{V}_6\hat{\boldsymbol{\mu}}_6).$$

It is clear from the mean model that x_1 and x_2 have interactions with x_6. Because x_6 (the deck face surface flatness) is a noise factor, robustness against it can be achieved by adjusting the two control factors x_1 and x_2. This cannot be understood from ordinary kriging predictor without performing additional sensitivity analysis (Chen, Jin, and Sudjianto, 2005).

2.3.2 Example 2: Piston Slap Noise Experiment

Piston slap is an unwanted engine noise resulting from piston secondary motion. A computer experiment was performed by varying 6 factors to minimize the noise. The factors were set clearance between the piston and the cylinder liner (x_1), location of peak pressure (x_2), skirt length (x_3), skirt profile (x_4), skirt ovality (x_5), and pin offset (x_6). The experimental design and the data are given in Table 4. More details of the experiment can be found in Hoffman, et al. (2003) and Li and Sudjianto (2005).

To apply the Bayesian variable selection, first we scale x_1, x_2, x_3, and x_6 to $[1.0, 3.0]$. For ordinary kriging, we obtain $\hat{\boldsymbol{\theta}} = (1.17, .01, .23, .01, .01, .71)$ and $CVPE(0) = 1.4511$. The $CVPE(m)$ and $R^2(m)$ are plotted in Fig. 10 based on the variables identified by the Bayesian variable selection technique. We see that the three variables x_{1l}, $x_{1l}x_{6l}$, and $x_{1q}x_{6l}$ give the minimum $CVPE(3) = 1.2777$. The corresponding $R^2(3) = .79$, which shows that the three variables alone explain about 79% of the variability in the data. Estimating $\boldsymbol{\theta}$ again, we obtain $\hat{\boldsymbol{\theta}} = (.01, .01, .09, 1.32, .01, .46)'$ and $CVPE(3) = 1.1168$. Thus we can expect about a 23% improvement in prediction

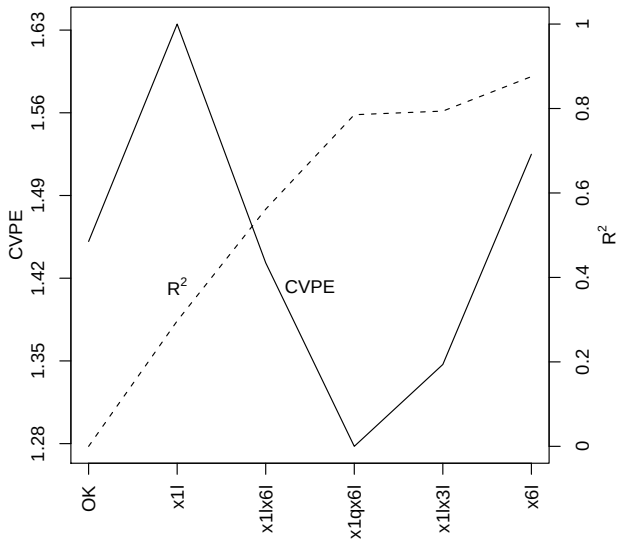

Figure 10: Plots of $CVPE(m)$ and $R^2(m)$ in Example 2

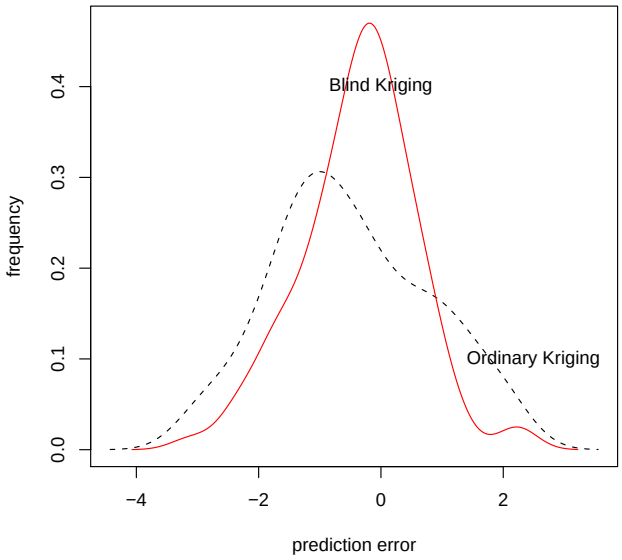

Figure 11: Density plot for the prediction errors in Example 2

Table 4: Example 2, Data for the piston slap noise experiment

Run	x_1	x_2	x_3	x_4	x_5	x_6	y
1	71	16.8	21	2	1	0.98	56.75
2	15	15.6	21.8	1	2	1.3	57.65
3	29	14.4	25	2	1	1.14	53.97
4	85	14.4	21.8	2	3	0.66	58.77
5	29	12	21	3	2	0.82	56.34
6	57	12	23.4	1	3	0.98	56.85
7	85	13.2	24.2	3	2	1.3	56.68
8	71	18	25	1	2	0.82	58.45
9	43	18	22.6	3	3	1.14	55.5
10	15	16.8	24.2	2	3	0.5	52.77
11	43	13.2	22.6	1	1	0.5	57.36
12	57	15.6	23.4	3	1	0.66	59.64

using blind kriging over ordinary kriging. The blind kriging predictor is given by

$$\hat{y}(\boldsymbol{x}) = 56.6 + 1.40x_{1l} - 1.12x_{1l}x_{6l} + .93x_{1q}x_{6l} + \hat{\boldsymbol{\psi}}(\boldsymbol{x})'\hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{y} - \boldsymbol{V}_3\hat{\boldsymbol{\mu}}_3).$$

We can see that in this example the $CVPE$ increased after the first step but then came down significantly after two more steps. This shows that we should not stop the procedure immediately when we observe an increase in $CVPE$. The procedure should be continued for a few more steps before choosing the value of m. Note that the R^2 plot is used only for interpretation and not for selecting the best m.

An additional 100 runs were performed for validating the results. The two densities of the prediction errors for ordinary kriging and blind kriging are shown in Fig. 11. It clearly shows that blind kriging gives a much better prediction. We can also calculate the root-mean squared prediction error (RMSPE) using

$$RMSPE = \sqrt{\frac{1}{100}\sum_{i=1}^{100}(y(\boldsymbol{x}_i) - \hat{y}(\boldsymbol{x}_i))^2}.$$

For ordinary kriging $RMSPE = 1.3626$ and for blind kriging $RMSPE = 1.0038$, which shows that the prediction error of blind kriging is smaller than that of ordinary kriging by about 26%.

There are several case studies reported in the literature where universal kriging is applied instead of ordinary kriging. Qian, et al. (2006) used universal kriging with all linear effects in the mean part of the model for the optimization in a material cellular design problem; see Sacks, Schiller, and Welch (1989) for other examples. In this example, we fitted a universal kriging model with linear effects for all of the factors. The universal kriging predictor is given by

$$\hat{y}(\boldsymbol{x}) = 55.3 + 1.02x_{1l} - .15x_{2l} - .96x_{3l} + .01x_{4l} - .45x_{5l} - .31x_{6l} + \hat{\boldsymbol{\psi}}(\boldsymbol{x})'\hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{y} - \boldsymbol{V}\hat{\boldsymbol{\mu}}),$$

with $\hat{\boldsymbol{\theta}} = (0.14, 0.01, 0.17, 0.01, 0.01, 0.09)$. The $RMSPE$ for the 100 validation runs is obtained as 1.5109, which is higher than both ordinary and blind kriging. The reason for this poor performance is that the mean part of the universal kriging model contains some unimportant effects ($R^2 = 25.4\%$). This shows the danger of using a universal kriging model without proper variable selection.

2.3.3 Example 3: Borehole Model

The following simple function for the flow rate through a borehole is used by many authors to compare different methods in computer experiments (see e.g., Morris, Mitchell, Ylvisaker (1993)):

$$y = \frac{2\pi T_u(H_u - H_l)}{ln(r/r_w)\left[1 + \frac{2LT_u}{ln(r/r_w)r_w^2K_w} + \frac{T_u}{T_l}\right]},$$

where the ranges of interest for the eight variables are $r_w : (0.05, 0.15)$, $r = (100, 50000)$, $T_u = (63070, 115600)$, $H_u = (990, 1110)$, $T_l = (63.1, 116)$, $H_l = (700, 820)$, $L = (1120, 1680)$, and $K_w = (9855, 12045)$. We re-scale the variables in $[1.0, 3.0]$ and denote them as $x_1, x_2, \cdots, x_8$. For convenience, we use the same 27-run experimental design in Table 3.

Using the Bayesian variable selection technique, we identified the linear effect of x_1 as the only important variable. The blind kriging predictor is given by

$$\hat{y}(\boldsymbol{x}) = 93.4 + 60.1x_{1l} + \hat{\boldsymbol{\psi}}(\boldsymbol{x})'\hat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{y} - \boldsymbol{V}_1\hat{\boldsymbol{\mu}}_1),$$

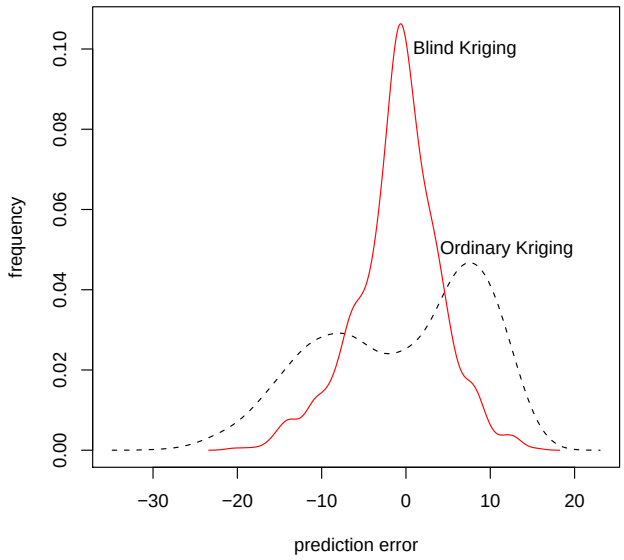

Figure 12: Density plot for the prediction errors in Example 3

with $\hat{\boldsymbol{\theta}} = (.31, .01, .01, .09, .01, .08, .07, .02)'$. We randomly generated 1,000 values within the experimental region and the prediction errors are plotted in Fig. 12. It shows remarkable improvement in prediction for blind kriging over ordinary kriging.

To check for the robustness against misspecification of correlation parameters, we repeated the calculations by varying $\boldsymbol{\theta}$. Let $\theta_1 = \cdots = \theta_8 = \theta$. Fig. 13 shows the plot of RMSPE values for different values of θ. We can see that the RMSPE values of blind kriging are almost half of those of ordinary kriging and have much less variation. This shows that blind kriging is more robust to misspecification in the correlation parameters than is ordinary kriging. This is a great advantage, because in practice it is difficult to obtain precise estimates of the correlation parameters.

We also tried universal kriging method for the borehole example. Two models are fitted, one with all linear terms and the other with all linear and quadratic terms. The RMSPE values for the 1,000 runs are given in Table 5. We can see that they are much higher than that of ordinary kriging and blind kriging. Thus, including unimportant variables in the mean part can actually deteriorate the performance.

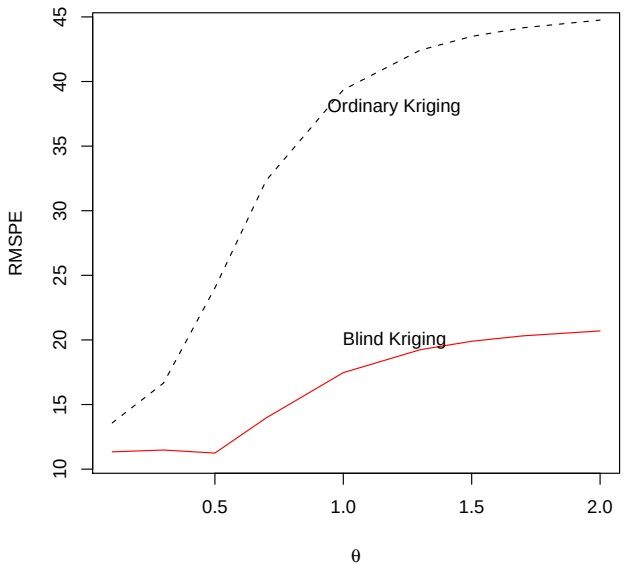

Figure 13: RMSPE values for different θ in Example 3

This clearly shows the importance of selecting variables carefully and the superiority of blind kriging over universal kriging.

Table 5: Comparison of different methods in Example 3

Method	m	RMSPE
Ordinary Kriging	0	9.7
Blind Kriging	1	5.5
Universal Kriging (linear)	8	11.3
Universal Kriging (linear and quadratic)	16	18.0

2.4 Conclusions

It is a common practice in the literature to use a constant mean for the kriging model. Although some recent studies point out the benefits of using more complex models for the mean (Martin and Simpson, 2006; Qian, et al., 2006), none of them have proposed a systematic methodology to obtain such models. In fact, the problem is much more complicated than merely using a complex model for the mean. Unnecessary variables

in the mean model can deteriorate the performance. Therefore only those variables that have a significant effect on the response should be used for the mean model. We showed that they can be identified using a Bayesian forward selection technique proposed in Joseph (2006b) and Joseph and Delaney (2007).

The Bayesian forward selection technique is directly related to kriging, which makes it attractive to use in blind kriging method. The most difficult step in this Bayesian technique is the estimation of correlation parameters. However, the estimates are readily available from ordinary kriging model and thus, the technique can be applied with no additional difficulty. Another advantage of the technique is that it incorporates the effect hierarchy and heredity principles through prior specification and thus, produces interpretable models.

We also note that a naive strategy of identifying important variables using a variable selection technique and then fitting the kriging part, in general will not work. This is because the performance of blind kriging is quite sensitive to the number of variables used in the mean part. Our approach computes the cross validation errors at each step of the Bayesian forward selection technique and selects the model with minimum error. It may happen that ordinary kriging itself is the optimal predictor, which cannot be detected in the naive strategy that applies a variable selection technique without considering the kriging part. Thus, we believe that the use of cross validation errors along with the Bayesian forward selection technique is critical for obtaining a good blind kriging predictor.

Several examples presented in the chapter demonstrate that substantial improvement in prediction can be achieved by using blind kriging. It is also shown that blind kriging predictor is simpler to interpret and is more robust to the misspecification in the correlation parameters than ordinary kriging predictor.

CHAPTER III

EXPERIMENTAL DESIGN AND ANALYSIS USING NESTED FACTORS WITH APPLICATIONS IN MACHINING

[1]

3.1 Introduction

Nested factors are those factors which exist only within the level of another factor. A factor within which other factors are nested is called a branching factor. For example, suppose we want to experiment with two surface preparation methods in printed circuit board (PCB) manufacturing: mechanical scrubbing and chemical treatment. Here, the surface preparation method is the branching factor. Mechanical scrubbing can be optimized by changing the pressure of the scrub and chemical treatment can be optimized by changing the micro-etch rate. The pressure and micro-etch rate are the nested factors. When designing an experiment, these two factors (pressure and micro-etch rate) will be collapsed into a single factor (i.e., they will be assigned to the same column in the design matrix). The physical meaning and levels of the nested factor depends on the corresponding level of the branching factor. Because nested factors can differ with respect to the level of branching factor, designing and analyzing experiments with such factors are not trivial.

Taguchi (1987) has proposed an innovative idea to design experiments with branching and nested factors. He called nested factors as pseudo-factors and the resulting

[1]The paper based on this chapter is under revision in *Technometrics*.

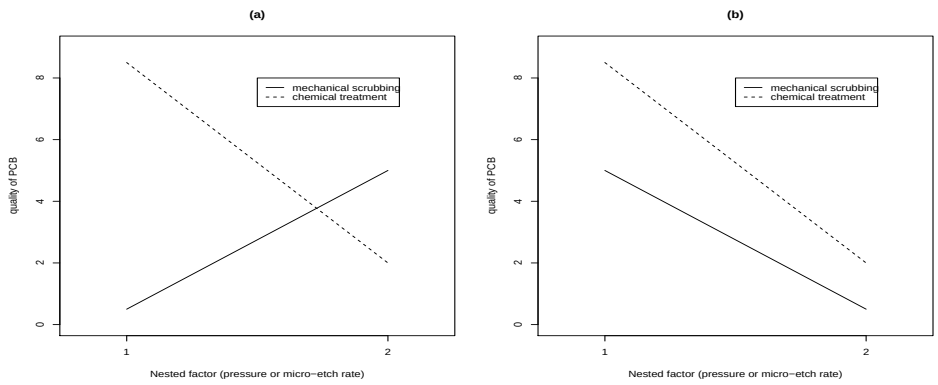

Figure 14: Illustration of branching-by-nested interaction. (a) when the effects are unknown and (b) when the effects are known.

designs as pseudo-factor designs. Phadke (1989) called the same as branching designs. The core idea was to carefully assign branching and nested factors to the columns of orthogonal arrays using linear graphs in such a way that their interactions are estimable. The interaction between branching and nested factors are important because the nested factors differ with respect to the levels of branching factors and thus their effects can change depending on the level of the branching factor. For example, suppose we choose two levels for the pressure and micro-etch rate in the PCB experiment. The quality of the PCB may increase with increase in pressure but may decrease with increase in micro-etch rate. This is shown in Figure 14(a). We can see strong interaction between the branching and nested factors. The interaction effect could be reduced if we knew the effects of the nested factors before the experiment. For example, if we interchange the two levels of pressure, then the interaction is not significant (see Figure 14 (b)). In general, the effects of the factors are not known before the experiment and therefore, we should design the experiments that are capable of estimating the branching-by-nested interactions. Although Taguchi's approach using orthogonal arrays and linear graphs are very intuitive, they are not general enough to apply to more complex situations such as the design of a computer experiment.

The designs we consider here are different from the so-called nested designs in the literature (see, e.g., Hicks and Turner, 1999; Montgomery, 2001). In nested designs, the factors are assumed to be similar (for example, different batches of material nested within different suppliers). Therefore, the branching-by-nested interactions can be safely assumed to be negligible, which is not the case in the present problem. Moreover, nested designs are crossed designs and therefore, the complication arising due to the aliasing of effects is not an issue. In contrast, the main focus here is to efficiently design highly fractionated experiments in the space of branching, nested, and other factors. Furthermore, in nested designs the nested factors are usually not the effects of interests and are treated as random effects or block effects, whereas our objective is to simultaneously identify the optimal settings of branching, nested, and other factors.

Our work is motivated by a computer experiment that involves branching and nested factors. The objective of the experiment is to optimize a turning process for hardened bearing steel with a cBN cutting tool (see Figure 15). This process is commonly referred to as hard turning and is of considerable interest to bearing manufacturers as a potential replacement for the grinding process. Since the material being machined is very hard (hardness in excess of 60 Rockwell C), the cutting tool is subjected to large forces, stresses, and temperatures during the operation. In practice, the cutting edge of the tool is shaped such that it can withstand the severe conditions. Two commonly employed cutting edge shapes, hone and chamfer, are shown in Figure 16. Note that Figure 16 represent the idealized view of the instantaneous cutting action in the cross section A-A indicated in Figure 15. These cutting edge shapes are intended to strengthen the cutting edge to bear the large tool stresses generated in cutting. The chamfer tool design can be changed using two factors: chamfer length and chamfer angle, whereas the hone design is fixed. In other words, the two factors length and angle are nested within the chamfer edge and there are no factors nested

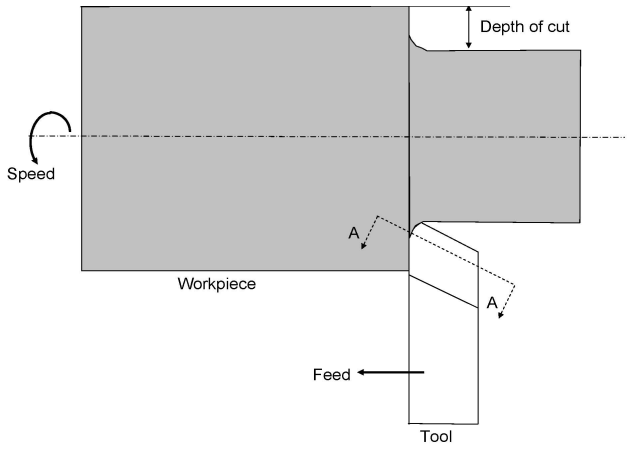

Figure 15: Schematic of turning process; A-A is a perpendicular section through the tool.

within hone edge. In our terminology, the tool edge is a branching factor. Thus, when the branching factor takes the level chamfer, there are two additional factors present in the experiment; but when the branching factor takes the level hone, there are no factors. There are a few other factors that are common to both of the tool edges such as the cutting edge radius, tool nose radius, and rake angle. The machining parameters such as cutting speed, feed, and depth of cut are also factors that do not depend on the type of tool edges. To distinguish them from the branching and nested factors, we call them as shared factors. All of the factors involved in this experiment and their allowable ranges are shown in Table 1. The experiments can be performed in computers using a commercially available finite element software AdvantEdge.

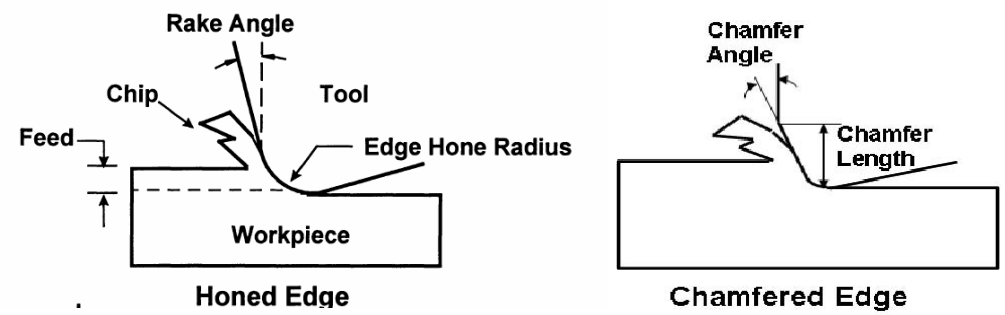

Figure 16: Illustration of hone and chamfer tool edges.

Table 6: Factors and their ranges in the hard turning experiment

Type of Factor	Notation	Factor	Ranges
Branching factor	z_1	Cutting edge shape	hone or chamfer
Nested factors	$v_1 \vert z_1$=chamfer	Angle (degree)	$17 \sim 20$
	$v_2 \vert z_1$=chamfer	Length (μm)	$115 \sim 140$
	$v_1 \vert z_1$=hone	None	None
	$v_2 \vert z_1$=hone	None	None
Shared factors	x_1	Cutting edge radius (μm)	$5 \sim 25$
	x_2	Rake angle (degree)	$-15 \sim -5$
	x_3	Tool nose radius (mm)	$0.4 \sim 1.6$
	x_4	Cutting speed (m/min)	$120 \sim 240$
	x_5	Feed (mm/rev)	$0.05 \sim 0.15$
	x_6	Depth of cut (mm)	$0.1 \sim 0.25$

Latin hypercube designs (LHDs) are commonly used in computer experiments (McKay, Beckman, and Conover, 1979). A desirable property of a LHD is its one-dimensional balance, i.e., when we project a N-point design onto any factor, there will be N different levels for that factor. Clearly this cannot be satisfied for branching and nested factors. The branching factor is usually a qualitative factor and therefore the number of levels of a branching factor is fixed (it does not depend on the number of runs N). Moreover, the nested factors are different for different levels of the branching factor. Therefore, we need a one-dimensional balance for the nested factors within each level of the branching factor. As an example, consider an experiment with one branching factor z_1, one nested factor v_1, and two shared factors x_1 and x_2. Suppose that the branching factor has two levels and that we want to do the experiment in eight runs. A possible design of experiment is shown in Table 2. We can see that the shared factors have the one-dimensional balance, because they take eight different levels in the experiment. The nested factor has a one-dimensional balance within each level of the branching factor. Note that v_1 represents two different factors, one when $z_1 = 1$ and the other when $z_1 = 2$. Therefore, $v_1 = 1$ in run 1 is not the same as $v_1 = 1$ in run 5. In the next section, we discuss some general strategies for designing

Table 7: An Example of Branching Latin hypercube design

run	z_1	v_1	x_1	x_2
1	1	1	4	1
2	1	2	3	8
3	1	3	8	5
4	1	4	2	3
5	2	1	7	2
6	2	2	1	6
7	2	3	6	7
8	2	4	5	4

such experiments using LHDs.

It is well known that not all LHDs are "good". Most of the research in this area has focused on finding "good" LHDs based on some optimal design criteria. See Iman and Conover (1982), Tang (1993), Owen (1994), Morris and Mitchell (1995), Tang (1998), Ye (1998), Ye, Li, and Sudjianto (2000), Jin, Chen, and Sudjianto (2005), and Joseph and Hung (2008). We need to extend those optimal design criteria for experiments with branching and nested factors. Take for example the case of maximin LHD proposed by Morris and Mitchell (1995). Here the optimal design criterion is to maximize the inter-site distance among the experimental points (runs). Now with the branching and nested factors, the notion of "distance" does not exist for all factors. Branching factors are qualitative and thus they can not be measured by distances. Moreover, for nested factors, the notion of "distance" exists only if the corresponding levels of the branching factor are the same. Another major aspect that makes the design of experiment different from the usual designs is the importance of the interaction between branching and nested factors. As noted before, these interactions are usually not negligible. Therefore, if any of the main effects is highly correlated with one of these interactions, then that main effect will be misspecified. Thus, the optimal design criteria should be modified to capture the branching-by-nested interaction effects.

Table 8: Illustration of the naive strategy

run	z_1	$v_1^{z_1} \cdots v_{m_1}^{z_1}$ x_1 $\cdots$ x_t
1	1	
$\vdots$	$\vdots$	LHD(n_0, $m_1 + t$)
$\vdots$	1	
$\vdots$	2	
$\vdots$	$\vdots$	LHD(n_0, $m_1 + t$)
$2n_0$	2	

The remaining of the chapter is organized as follows. In Section 2, we discuss some general strategies for designing experiments with branching and nested factors and introduce the concept of branching LHD. In Section 3, we discuss three different criteria for finding an optimal branching LHD. The analysis of experiments with branching and nested factors is discussed in Section 4. We illustrate the proposed methods using the hard turning experiment in Section 5. Some concluding remarks and future research directions are given in Section 6.

3.2 Branching Latin Hypercube Designs

In general, an LHD with N runs and p factors, denoted by LHD(N, p), can be generated using a random permutation of $\{1, 2, \cdots, N\}$ for each factor. However, as discussed before, this cannot be done if the experiment involves branching and nested factors.

Consider a simple case where the branching factor z_1 has only two levels and m_1 factors are nested within each level of the branching factor. Thus, there are m_1 nested factors ($v_1^{z_1}$, $\cdots$, $v_{m_1}^{z_1}$), where each of the nested factor stands for two different factors depending on the two levels of the branching factor. In addition, there are t more shared quantitative factors. Now we discuss some strategies for constructing LHDs with branching and nested factors.

A naive strategy is to choose an LHD for the nested and shared factors and repeat it under each level of the branching factor. That is, first we choose an LHD(n_0, m_1+t) that can accommodate the nested and shared factors. This will then be repeated for the two levels of the branching factor. The resulting design is shown in Table 8. This is easy to construct. Moreover, we can easily choose optimal LHDs for the nested and shared factors using the existing methods. In addition, because all of the combinations of nested and shared factors are repeated at each level of the branching factor, we can estimate the interactions involving branching factor. However, there are two drawbacks. One is that if we project the design matrix onto one of those shared factors ($x_1, \cdots, x_t$), there are some replications. Hence, the design points are not spread out as uniformly as they could be. The other drawback is that, the run size of these designs can be quite large.

The foregoing problems with the naive strategy can be easily overcome by using one LHD for all of the shared factors. As a result, the design points are spread out more uniformly in the experimental region and the run size required is comparably smaller. As an example, for $m_1 = 3$ and $t = 5$, the run size of the naive approach ($2n_0$ in Table 8) should be at least 16 (because $n_0 \geq 8$). It can be reduced to 6 ($2n_1 \geq 6$) by the new design illustrated in Table 9. Following the terminology used by Phadke (1989), we name a design with this structure a *branching Latin hypercube design* (BLHD).

Now consider a more general case with q branching factors denoted by $\boldsymbol{z} = (z_1, \cdots, z_q)$. Assume that all of them are qualitative by nature. For each branching factor z_u, there are k_u levels and under each of these different levels, there are m_u nested factors. Note that in general, the number of factors nested under each level of a branching factor can be different. For example, in the hard turning experiment there are two factors nested under chamfer but none under hone. However, for notational simplicity, we assume the number of nested factors to be the same (for a given

Table 9: Branching Latin hypercube design with one branching factor

run	z_1	$v_1^{z_1} \cdots v_{m_1}^{z_1}$	$x_1 \ \cdots \ x_t$
1	1		
⋮	⋮	LHD(n_1, m_1)	
⋮	1		LHD($2n_1$, t)
⋮	2		
⋮	⋮	LHD(n_1, m_1)	
$2n_1$	2		

branching factor) and develop the construction of BLHD. Later we explain how it can be extended to deal with unequal number of nested factors.

Denote the nested factors by $\boldsymbol{v}^{z_u} = (v_1^{z_u}, \cdots, v_{m_u}^{z_u})'$, $1 \leq u \leq q$. Again note that, each nested factor corresponds to different factors depending on the branching factor and its level. That is why we use a superscript to denote the branching factor level. In addition to the branching and nested factors, there are t shared quantitative factors $\boldsymbol{x} = (x_1, \cdots, x_t)'$. Let $\boldsymbol{v} = ((\boldsymbol{v}^{z_1})', \cdots, (\boldsymbol{v}^{z_q})')'$, and $\boldsymbol{w} = (\boldsymbol{x}', \boldsymbol{z}', \boldsymbol{v}')'$ represents all of the p factors involved in the experiment, where $p = t + q + \sum_{u=1}^{q} m_u$. A N-run BLHD can then be represented by $\boldsymbol{W} = (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_N)'$. In general, it consists of three parts. The first part is a design for branching factors. Because branching factors are qualitative factors, we can choose an orthogonal array of appropriate size depending on the number of levels of each branching factor. The second part consists of LHDs for the nested factors. Choose LHD(n_u, m_u) for the m_u nested factors under the branching factor z_u, $u = 1, 2, \cdots, q$. The third part is a LHD(N, t) for all of the shared factors $\boldsymbol{x}$. If there are k_u levels for branching factor z_u, where $1 \leq u \leq q$, it is clear that $N = k_1 n_1 = k_2 n_2 = \cdots = k_q n_q$. Thus, we have one orthogonal array for the branching factors, q LHDs for the nested factors, and one LHD for the shared factors. These designs can be assembled to obtain a BLHD.

As an example, consider the case with two branching factors (z_1 and z_2) each at

Table 10: Branching Latin hypercube design with two branching factors

run	z_1	$v_1^{z_1} \cdots v_{m_1}^{z_1}$	z_2	$v_1^{z_2} \cdots v_{m_2}^{z_2}$	$x_1 \cdots x_t$
1	1	LHD(n_1, m_1)	1	LHD(n_2, m_2) first half	
⋮	1		2	LHD(n_2, m_2) first half	LHD(N, t)
⋮	2	LHD(n_1, m_1)	1	LHD(n_2, m_2) second half	
N	2		2	LHD(n_2, m_2) second half	

two levels. There are m_1 nested factors under z_1 and m_2 nested factors under z_2. Furthermore, there are t shared factors. Table 10 illustrates a N-run BLHD for this example. The first part is a 4-run orthogonal array for those two branching factors. For the second part, we choose LHD(n_1, m_1) for the nested factors under z_1. Similarly, LHD(n_2, m_2) is chosen for the nested factors under z_2. This LHD is divided into two halves and distributed among the two levels of z_2 as shown in the Table. The third part consists of a LHD(N, t) for the t shared factors.

As in the case with LHDs, not all BLHDs are good. We need to use some optimal design criteria to choose the best BLHD. This is discussed in the next section.

3.3 Optimal Branching Latin Hypercube Designs

As discussed in the introduction, there are several approaches for finding a good LHD. We may think it is enough to use one of those approaches to generate $q + 1$ optimal LHDs for the nested and shared factors and assemble them to obtain the BLHD. However, such an assembly of optimal LHDs may not lead to an optimal BLHD. Moreover, we need to make sure that in a BLHD, the correlation between the branching-by-nested interaction and any other main effect is small. In this section, we propose three optimal design criteria for finding good BLHDs.

3.3.1 Maximin BLHD

Morris and Mitchell (1995) proposed to find LHDs that maximize the minimum inter-site distance. Let $\boldsymbol{g}$ and $\boldsymbol{h}$ be two design points (or sites or runs). Consider the distance measure $d(\boldsymbol{g}, \boldsymbol{h}) = \{\sum_{j=1}^{p} |g_j - h_j|^{\varsigma}\}^{1/\varsigma}$, in which $\varsigma = 1$ and $\varsigma = 2$ correspond to the rectangular and Euclidean distance, respectively. For simplicity, rectangular distance ($\varsigma = 1$) is considered for the rest of the chapter. For a given LHD, define a distance list $(D_1, D_2, \cdots, D_M)$ in which the elements are the distinct values of inter-site distances, sorted from the smallest to the largest. Let J_i be the number of pairs of design points in the design separated by D_i. Then a design is called a maximin design if it sequentially maximizes D_i's and minimizes J_i's in the following order: $D_1, J_1, D_2, J_2, \cdots, D_M, J_M$. A scalar-valued function which can be used to rank competing designs in such a way that the maximin design receives the highest ranking is given by

$$\phi_\lambda = (\sum_{i=1}^{M} J_i D_i^{-\lambda})^{1/\lambda} = \left(\sum_{\boldsymbol{g} \neq \boldsymbol{h}} d(\boldsymbol{g}, \boldsymbol{h})^{-\lambda} \right)^{1/\lambda}, \tag{1}$$

where λ is a positive integer.

Extension of the maximin criterion to BLHDs is not straightforward. Different from LHDs where all factors can be measured by distances, BLHDs have branching factors which have no notion of distance and nested factors where definition of distance depend on the corresponding branching factors. Because of different roles of factors, instead of calculating all the pairwise distances over all factors, we need to consider branching and nested factors separately from those of shared factors. First note that for BLHDs, not all factors are divided into same number of levels as that in LHDs: there are k_u levels for branching factor z_u, n_u levels for nested factors $\boldsymbol{v}^{z_u}$, and N levels for $\boldsymbol{x}$. Therefore, before calculating the distances, the design matrix should be scaled to $(-1, 1)$.

Start with a simple case where $q = 1$ and $m_1 = 1$. Thus, there are $t + 2$ factors in

the experiment. Assume that the last two factors are branching factor z_1 and nested factor $v_1^{z_1}$, respectively. We need to define two types of inter-site distances. The first type of distance focuses on all of the shared factors. It is the distance projection onto the t-dimensional space ($\boldsymbol{x}$), which can be defined by $d_x(\boldsymbol{g}, \boldsymbol{h}) = \sum_{j=1}^{t} |g_j - h_j|$, where $\boldsymbol{g} = (g_1, \cdots g_{t+2})$ and $\boldsymbol{h} = (h_1, \cdots, h_{t+2})$ are $(t+2)$-dimensional design points. There are a total of $\binom{N}{2}$ distances. The second type of distance takes into account of the branching and nested factors by considering distances within each level of branching factors. The objective here is to spread out the design points for each level of branching factors. To do so, for each level $z_{1,i}$ of the branching factor, where $1 \leq i \leq k_1$, distances are calculated based on $\boldsymbol{x}$ and $v_1^{z_1}$. Define the second type of distance by $d_B(\boldsymbol{g}, \boldsymbol{h}) = \sum_{l=1}^{t} |g_l - h_l| + |g_{t+2} - h_{t+2}|$. One can easily obtain $d_B(\boldsymbol{g}, \boldsymbol{h}) = d_{v_1}(\boldsymbol{g}, \boldsymbol{h}) + d_x(\boldsymbol{g}, \boldsymbol{h})$, where $d_{v_1}(\boldsymbol{g}, \boldsymbol{h}) = |g_{t+2} - h_{t+2}|$. The second type of inter-site distances are calculated only for those within the same level of the branching factor ($g_{t+1} = h_{t+1} = z_{1,i}$, for some $1 \leq i \leq k_1$) and thus there are $\binom{n_1}{2} k_1$ of them.

Note that, the first type of distance measure consists of t dimensions, while the second consists of $t+1$ dimensions. After standardizing with respect to their dimensions, the maximin distance criterion can be extended to BLHDs by defining a distance list based on these $\binom{N}{2} + \binom{n_1}{2} k_1$ standardized inter-site distances. Furthermore, as in (1), the scalar-valued function can be defined as

$$\phi_\lambda = \left(\sum_{\boldsymbol{g} \neq \boldsymbol{h}} \left[\frac{t}{d_x(\boldsymbol{g}, \boldsymbol{h})} \right]^\lambda + \sum_{i=1}^{k_1} \sum_{g_{t+1} = h_{t+1} = z_{1,i}} \left[\frac{1+t}{d_{v_1}(\boldsymbol{g}, \boldsymbol{h}) + d_x(\boldsymbol{g}, \boldsymbol{h})} \right]^\lambda \right)^{1/\lambda}, \qquad (2)$$

where $\sum_{g_{t+1} = h_{t+1} = z_{1,i}} d_{v_1}(\boldsymbol{g}, \boldsymbol{h})$ is the sum of $\binom{n_1}{2}$ pairwise distances in which $\boldsymbol{g}$ and $\boldsymbol{h}$ have the same level of branching factor and $\sum_{\boldsymbol{g} \neq \boldsymbol{h}} d_x(\boldsymbol{g}, \boldsymbol{h}) = \sum_{i=1}^{k_1} \sum_{g_{t+1} = h_{t+1} = z_{1,i}} d_x(\boldsymbol{g}, \boldsymbol{h}) + \sum_{g_{t+1} \neq h_{t+1}} d_x(\boldsymbol{g}, \boldsymbol{h})$.

To illustrate this idea, consider the simple example in Table 7. Assume that the branching factor z_1 is qualitative. The optimal design found by the modified maximin criterion (2) (with $\lambda = 15$) is $x_1 = \{1, 5, 6, 4, 7, 3, 2, 8\}$, $x_2 = \{4, 8, 1, 5, 6, 2, 7, 3\}$, and

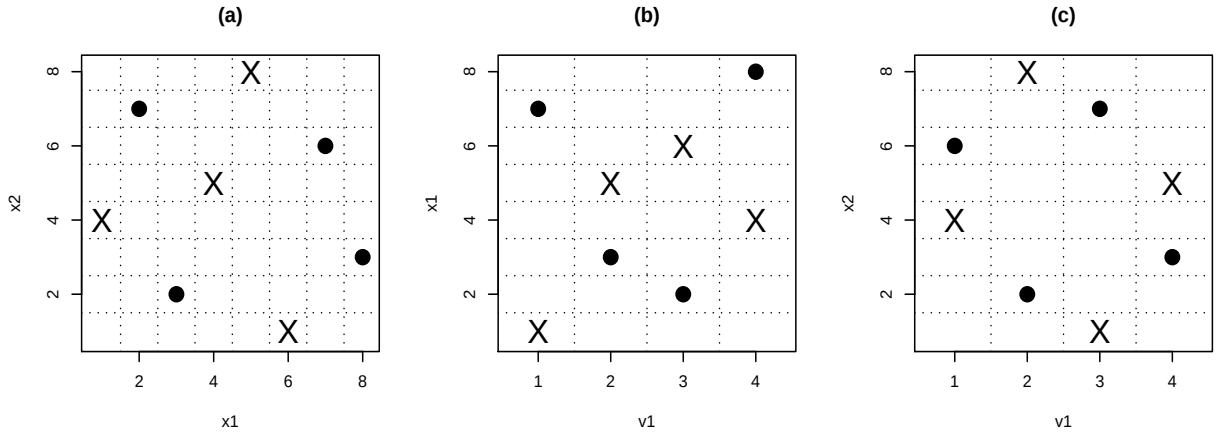

Figure 17: Maximin BLHD. "X" stands for $z_1 = 1$ and solid points stands for $z_1 = 2$.

z_1 and v_1 remain the same as in Table 7. This maximin BLHD is plotted in Figure 17, where "X" represent those design points with $z_1 = 1$ and solid points stands for those with $z_1 = 2$. The first part in (2) tries to maximize the inter-site distances in the space of x_1 and x_2 (Figure 17(a)), in which the "X" and solid points are not distinguished. Whereas the second part in (2) tries to maximize the inter-site distances in the space of x_1, x_2, and v_1 (Figures 17(b) and (c)). Moreover, because these distances are calculated only within the same level of the branching factor, the inter-site distances among the "X" points and among the solid points are maximized.

If we were to use only the first part in (2) as the criterion, then the optimal design would be space-filling only over the shared factors. The design points can be quite structured with respect to the branching and nested factors. This can be seen in Figure 18. It is clear that although the design points are evenly spread out in the (x_1, x_2) space (Figure 18(a)), experiments for $z_1 = 1$ concentrate on lower level of x_1 and experiments for $z_1 = 2$ concentrate on higher level of x_1. Furthermore, the nested factor v_1 is highly correlated with x_1. This clearly shows the importance of the new criterion in (2). We should also point out that the two designs for the shared factors (Figure 17(a) and Figure 18(a)) are isomorphic. However, these two isomorphic designs are not equally good when we include the branching factor. The

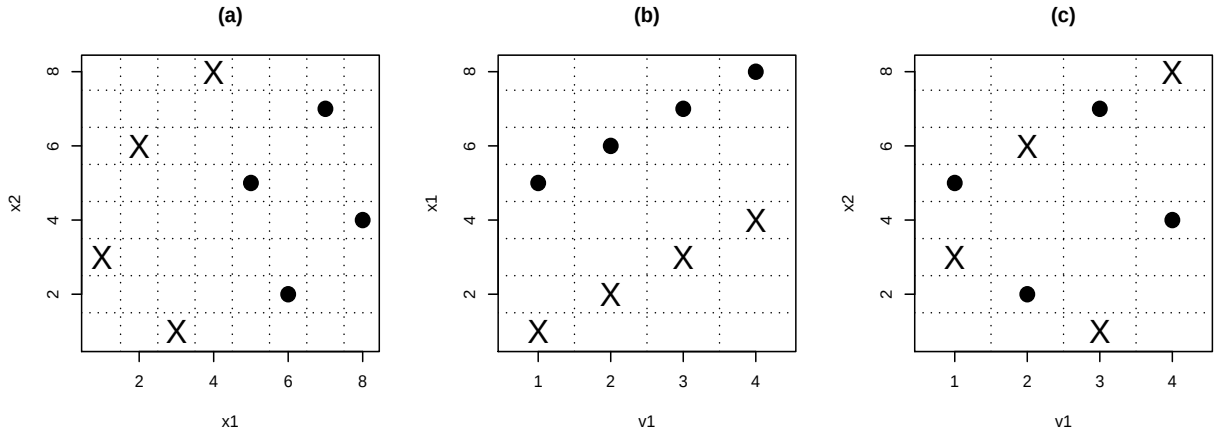

Figure 18: Maximin for shared factors

new criterion could clearly identify that Figure 17(a) gives a better design than Figure 18(a).

Now consider the general case with q branching factors $\boldsymbol{z}$ and the corresponding nested factors $\boldsymbol{v}$. Assume that for each branching factor z_u, there are k_u levels denoted by $z_{u,i}$, $1 \leq i \leq k_u$. Let $\boldsymbol{g}$ and $\boldsymbol{h}$ are design points with p dimensions. Let $\delta_u = (t + \sum_{l=1}^{u}(m_{l-1}+1)+1)$, such that the δ_uth factor is the uth branching factor, and the $[\delta_u+1]$th to $[\delta_u+m_u]$th factors are the corresponding nested factors. As an extension of (2), the maximin criterion for BLHDs can be written as

$$\phi_\lambda = \left(\sum_{\boldsymbol{g}\neq\boldsymbol{h}} \left[\frac{t}{d_x(\boldsymbol{g},\boldsymbol{h})}\right]^\lambda + \sum_{u=1}^{q}\sum_{i=1}^{k_u} \sum_{g_{\delta_u}=h_{\delta_u}=z_{u,i}} \left[\frac{m_u+t}{d_{v_u}(\boldsymbol{g},\boldsymbol{h})+d_x(\boldsymbol{g},\boldsymbol{h})}\right]^\lambda \right)^{1/\lambda}, \qquad (3)$$

where $d_x(\boldsymbol{g},\boldsymbol{h}) = \sum_{j=1}^{t}|g_j - h_j|$ and the distance measure for the u-th branching factor is denoted by $d_{v_u}(\boldsymbol{g},\boldsymbol{h}) = \sum_{l=\delta_u+1}^{\delta_u+m_u}|g_l - h_l|$. This criterion is general and includes some interesting special cases.

Case 1: If $q = 0$, and $m_u = 0$ for all u, then this would lead to the standard LHD(N,t). Up to a constant, the maximin criterion (3) would be the same as (1) proposed by Morris and Mitchell (1995).

Case 2: If there is no nested factor corresponding to the branching factors, that is $m_u = 0$, for all u, one can think about this as an experiment with q qualitative

factors $\boldsymbol{z}$ and t quantitative factors $\boldsymbol{x}$. As a special case of (3), the maximin criterion for experimental design with quantitative and qualitative factors can be written as

$$\phi_\lambda = \left(\sum_{\boldsymbol{g} \neq \boldsymbol{h}} \left[\frac{t}{d_x(\boldsymbol{g}, \boldsymbol{h})} \right]^\lambda + \sum_{u=1}^{q} \sum_{i=1}^{k_u} \sum_{g_{\delta_u} = h_{\delta_u} = z_{u,i}} \left[\frac{t}{d_x(\boldsymbol{g}, \boldsymbol{h})} \right]^\lambda \right)^{1/\lambda} \tag{4}$$

Case 3: Another special case is that when $t = 0$. In this situation, experiments consist branching factors and nested factors but no shared factors. Because $d_x(\boldsymbol{g}, \boldsymbol{h}) = 0$, for all $\boldsymbol{g}$ and $\boldsymbol{h}$, (3) can be simplified to

$$\phi_\lambda = \left(\sum_{u=1}^{q} \sum_{i=1}^{k_u} \sum_{g_{\delta_u} = h_{\delta_u} = z_{u,i}} \left[\frac{m_u}{d_v(\boldsymbol{g}, \boldsymbol{h})} \right]^\lambda \right)^{1/\lambda}.$$

3.3.2 Minimum Correlation BLHD

Apart from space-filling, another important issue regarding experimental designs is how to construct them such that the significant factors can be correctly identified. For LHDs, it can be achieved by minimizing the pairwise correlation among factors (Iman and Conover, 1982; Owen, 1994; Tang, 1998). Owen (1994) proposed a performance measure ρ^2 for evaluating the goodness of an LHD with respect to pairwise correlations. For LHD(N, t),

$$\rho^2 = \frac{\sum_{i=2}^{t} \sum_{j=1}^{i-1} \rho_{ij}^2}{t(t-1)/2}, \tag{5}$$

where ρ_{ij} is the linear correlation between columns i and j.

Different from LHDs, orthogonality among main effects is not enough in BLHDs. In BLHDs, it is equally important to consider the branching-by-nested interactions. Therefore, we propose a modified correlation criterion, which minimizes the correlations among the main effects of all factors as well as those between main effect of a shared factor and a branching-by-nested interaction effect. To do so, we first enlarge the BLHDs by including two-factor interactions which represent the branching-by-nested interaction. There are m_u such interactions for each branching factor z_u,

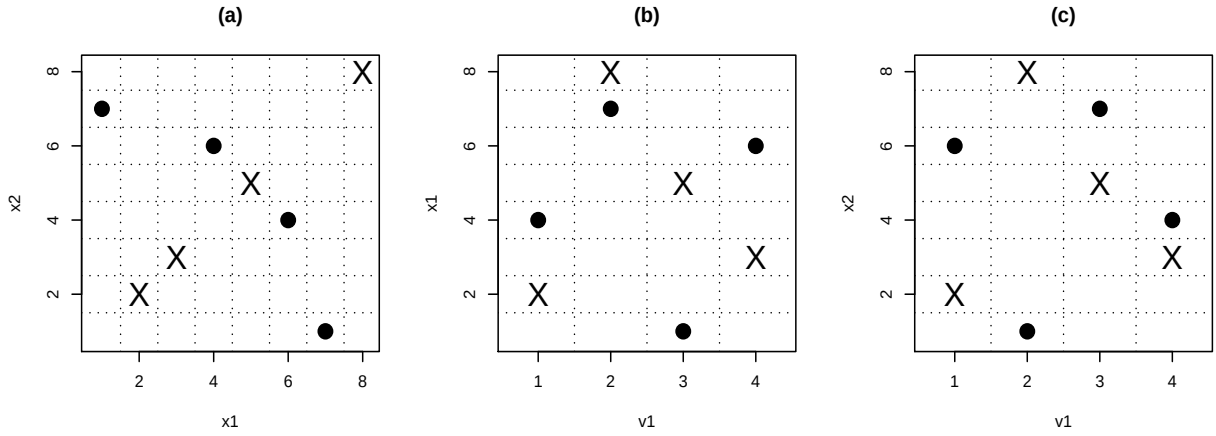

Figure 19: Minimum correlation BLHD

where m_u is the number of nested factors. Therefore, the total number of branching-by-nested interactions would be s, where $s = \sum_{u=1}^{q} m_u$. Thus, the new criterion for BLHDs is given by

$$\rho^2 = \frac{\sum_{i=2}^{p}\sum_{j=1}^{i-1}\rho_{ij}^2 + \sum_{i=1}^{t}\sum_{j=1}^{s}\tilde{\rho}_{ij}^2}{(p(p-1)/2) + st}, \tag{6}$$

where ρ_{ij}^2 is the linear correlation between columns i and j in the design and $\tilde{\rho}_{ij}^2$ is the linear correlation between x_i and jth branching-by-nested interaction.

Consider the same example used in Table 7. The optimal design that minimizes ρ^2 in (6) is $x_1 = (2, 8, 5, 3, 4, 7, 1, 6)$, $x_2 = (2, 8, 5, 3, 6, 1, 7, 4)$, and z_1 and v_1 remain the same as in Table 7. The design is plotted in Figure 19.

3.3.3 Orthogonal-Maximin BLHD

Maximizing minimum inter-site distances does not ensure minimizing pairwise correlations and vice-versa. Therefore, Joseph and Hung (2008) proposed a multi-objective criterion for LHD that combines the maximin distance and the minimum correlation criteria. This criterion becomes even more important in the case of BLHDs, because it is important to ensure small correlations between the shared factors and the branching-by-nested interactions besides enuring good space-filling properties. To extend the result in Joseph and Hung (2008) to BLHD, we should scale ϕ_λ and ρ^2 to the

same range so that some meaningful weights can be assigned in the multi-objective function. The following result gives the lower and upper bounds for ϕ_λ, which can be used for scaling it to $[0, 1]$. Here we only consider the case of a single branching factor. The result can be extended to include more than one branching factor, but the expressions become more complicated.

PROPOSITION 2. *If there is only one branching factor, then*

$$\phi_{\lambda,L} \leq \phi_\lambda \leq \phi_{\lambda,U},$$

where

$$\phi_{\lambda,L} = 3\Bigg[\sum_{i=1}^{k_1} \sum_{g_{\delta_1}=h_{\delta_1}=z_{1,i}} 2^{\frac{1}{p+1}}(m_1+t)^{\frac{p}{p+1}} + \sum_{g_{\delta_1}\neq h_{\delta_1}} t^{\frac{p}{p+1}}\Bigg]^{\frac{(\lambda+1)}{\lambda}} \Bigg[t(N^2-1)+\sum_{i=1}^{k_1} m_1(n_1^2-1)\Bigg]^{-1},$$

and

$$\phi_{\lambda,U} = \frac{N}{2}\left[\sum_{i=1}^{k_1}\sum_{j=1}^{n_1-1} \frac{(n_1-j)(t+m_1)^\lambda}{j^\lambda(t+k_1m_1)^\lambda} + \sum_{j=1}^{N-1}\frac{N-j}{j^\lambda}\right]^{1/\lambda}.$$

Thus, the multi-objective criterion is to minimize

$$\psi_\lambda = w\rho^2 + (1-w)\frac{\phi_\lambda - \phi_{\lambda,L}}{\phi_{\lambda,U} - \phi_{\lambda,L}}. \tag{7}$$

We usually take $w = .5$ and call the design that minimizes this criterion as orthogonal-maximin BLHD.

The design matrix in Table 7 is an orthogonal-maximin BLHD. It is plotted in Figure 20. Comparisons of the optimal designs found by the forgoing three criteria are provided in Table 11. It can be seen that the correlation between the shared factor and the branching-by-nested interaction (denoted by INT) is 0 in the case of minimum correlation design. However, the points are much closer compared to the maximin BLHD. The orthogonal-maximin BLHD provides a good compromise between them.

Because of the combinatorial nature of the optimization problem, finding the optimal BLHD for large dimensions is a difficult task. Several methods such as

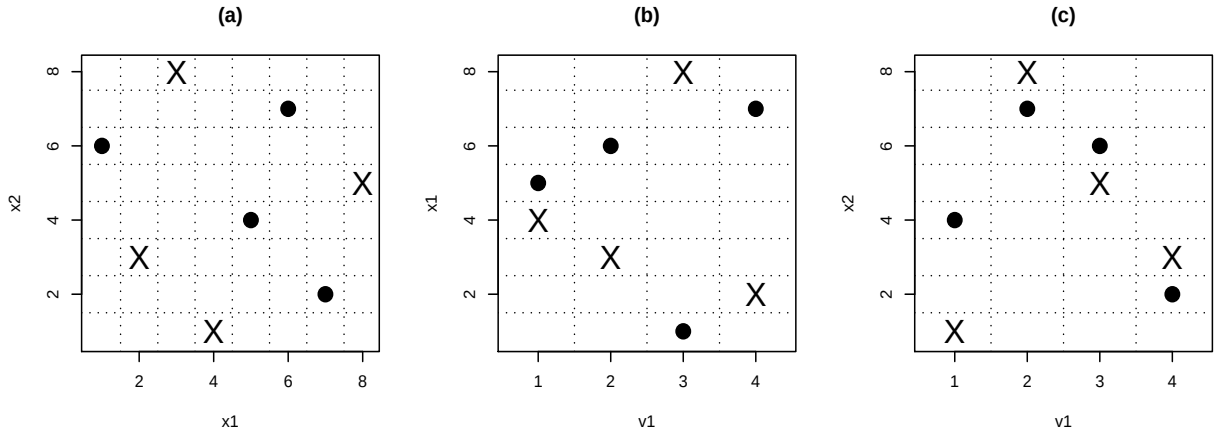

Figure 20: Orthogonal-maximin BLHD

Table 11: Comparison of different BLHDs

	Maximin Distance	Minimum Correlation	Orthogonal-Maximin
ϕ_λ	2.36	4.35	2.37
$D_1(J_1)$	$\frac{1}{2}$ (12)	$\frac{1}{4}$ (3)	$\frac{1}{2}$ (13)
ρ^2	0.0287415	0.000283	0.01445
(cor(x_1, INT), cor(x_2, INT))	(0.195,0)	(0,0)	(-0.098,0)

simulated annealing (Morris and Mitchell 1995), columnwise-pairwise algorithm (Ye, Li, and Sudjianto 2000), enhanced stochastic evolutionary algorithm (Jin, Chen, and Sudjianto 2005), and the modified simulated annealing (Joseph and Hung 2008) are proposed in the literature for finding the optimal LHD. These methods can be easily adapted for finding the optimal BLHD as well. A C++ code based on Joseph and Hung's algorithm is available from the authors upon request.

3.4 Kriging with Branching and Nested Factors

In this section we explain how branching and nested factors can be incorporated in kriging. Although similar extensions can be made on other methods such as linear regression, we focus here on kriging because of its popularity in the analysis of computer experiments (Sacks et al. 1989). The ordinary kriging model is given by

$Y(\boldsymbol{w}) = \mu + Z(\boldsymbol{w})$, where $Z(\boldsymbol{w})$ is a weak stationary stochastic process with mean 0 and covariance function $\sigma^2\psi$ and $\boldsymbol{w} \in \mathbb{R}^p$. The correlation function is defined as $\text{cor}\{Y(\boldsymbol{w}_1), Y(\boldsymbol{w}_2)\} = \psi(\boldsymbol{w}_1, \boldsymbol{w}_2)$. Usually, a product correlation structure is assumed for the correlation function. Consider the example in Table 7. The correlation function between two points $\boldsymbol{w}_1 = (x_{11}, x_{12}, z_{11}, v_{11}^{z_{11}})$ and $\boldsymbol{w}_2 = (x_{21}, x_{22}, z_{21}, v_{21}^{z_{21}})$ can be described as a product of correlation functions of each factor (ψ_i). In the usual cases, a common correlation function is chosen for each factor. However, this cannot be done in the present problem because of the different types of factors.

For the shared factors, a Gaussian correlation function may be used (Santner et al. 2003):

$$\psi_i(x_{1i}, x_{2i}) = \exp\{-\alpha_i(x_{1i} - x_{2i})^2\},$$

whereas for a branching factor, an isotropic correlation function may be used (Joseph and Delaney 2007):

$$\xi_1(z_{11}, z_{21}) = \exp\Big\{ - \theta_1 I_{[z_{11} \neq z_{21}]} \Big\}.$$

Here α_i and θ_1 are correlation parameters and I_A is an indicator function which takes value 1 when A is true and 0 otherwise.

A new correlation function needs to be developed for nested factors. Assume that they are quantitative factors. We cannot use the Gaussian correlation function here because a nested factor represents different factors depending on the level of the branching factor. Therefore, it is not reasonable to use one correlation parameter for a given nested factor. Instead, they should be different depending on the level of branching factor. For the example in Table 7, the correlation function for $v_1^{z_1}$ can be defined as following. If two points have the same level in the branching factor, for example $z_{11} = z_{21} =$ "1", then the correlation function will be $\exp\{-\gamma_1(v_{11}^{z_{11}} - v_{21}^{z_{21}})^2\}$. Similarly, if $z_{11} = z_{21}$="2", the correlation function will be $\exp\{-\gamma_2(v_{11}^{z_{11}} - v_{21}^{z_{21}})^2\}$. If the two points do not have the same level in the branching factor (i.e., $z_{11} \neq z_{21}$), then correlation should be determined by the branching factor not the nested factor.

Hence, in this case, the correlation function for nested factor will be equal to 1. Succinctly, the correlation function for the nested factor can be defined as

$$\varpi_1(v_{11}^{z_{11}}, v_{21}^{z_{21}}) = \exp\Big\{ -\sum_{j=1}^{2} \gamma_j (v_{11}^{z_{11}} - v_{21}^{z_{21}})^2 I_{[z_{11}=z_{21}=j]} \Big\}. \tag{8}$$

We can easily extend this to a more general situation. Let there are q branching factors $z_1, \cdots, z_q$. For each branching factor z_u, there are m_u nested factors. Assume $\boldsymbol{x}_1 = (x_{11}, \cdots x_{1t})'$, $\boldsymbol{z}_1 = (z_{11}, \cdots, z_{1q})'$, $\boldsymbol{v}_1^{z_u} = (v_{11}^{z_u}, \cdots, v_{1\,m_u}^{z_u})'$, and $\boldsymbol{v}_1 = ((\boldsymbol{v}_1^{z_1})', \cdots, (\boldsymbol{v}_1^{z_q})')$. Similarly, $\boldsymbol{x}_2 = (x_{21}, \cdots x_{2t})'$, $\boldsymbol{z}_2 = (z_{21}, \cdots, z_{2q})'$, $\boldsymbol{v}_2^{z_u} = (v_{21}^{z_u}, \cdots, v_{2m_u}^{z_u})'$, and $\boldsymbol{v}_2 = ((\boldsymbol{v}_2^{z_1})', \cdots, (\boldsymbol{v}_2^{z_q})')$. Given any two design points $\boldsymbol{w}_1 = (\boldsymbol{x}_1, \boldsymbol{z}_1, \boldsymbol{v}_1)$ and $\boldsymbol{w}_2 = (\boldsymbol{x}_2, \boldsymbol{z}_2, \boldsymbol{v}_2)$, the correlation function can be written as

$$\mathrm{cor}(Y(\boldsymbol{w}_1), Y(\boldsymbol{w}_2)) = \Big(\prod_{i=1}^{t} \psi_i(\boldsymbol{x}_1, \boldsymbol{x}_2) \Big) \prod_{u=1}^{q} \Big[\xi_u(\boldsymbol{z}_1, \boldsymbol{z}_2) \Big(\prod_{j=1}^{m_u} \varpi_u(v_{1j}^{z_{1u}}, v_{2j}^{z_{2u}}) \Big) \Big], \tag{9}$$

where $\psi_i(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\{-\alpha_i (x_{1i} - x_{2i})^2\}$ is the correlation function for the shared factors, $\xi_u(\boldsymbol{z}_1, \boldsymbol{z}_2) = \exp\Big\{ -\theta_u I_{[z_{1u} \neq z_{2u}]} \Big\}$ is the correlation function for the branching factors, and

$$\varpi_u(v_{1i}^{z_{1u}}, v_{2i}^{z_{2u}}) = \exp\Big\{ -\sum_{j=1}^{k_u} \gamma_{uij} (v_{1i}^{z_{1u}} - v_{2i}^{z_{2u}})^2 I_{[z_{1u}=z_{2u}=z_{u,j}]} \Big\}, \tag{10}$$

is the correlation function for the nested factors. Note that $z_{u,j}$, $1 \leq j \leq k_u$ are the k_u levels for each branching factor z_u. Thus, we obtain

$$\begin{aligned} \mathrm{cor}(Y(\boldsymbol{w}_1), Y(\boldsymbol{w}_2)) = \; & \exp\Big\{ -\textstyle\sum_{i=1}^{t} \alpha_i (x_{1i} - x_{2i})^2 \\ & - \textstyle\sum_{u=1}^{q} \Big[\theta_u I_{[z_{1u} \neq z_{2u}]} + \sum_{i=1}^{m_u} \sum_{j=1}^{k_u} \gamma_{uij} (v_{1i}^{z_{1u}} - v_{2i}^{z_{2u}})^2 I_{[z_{1u}=z_{2u}=z_{u,j}]} \Big] \Big\}. \end{aligned} \tag{11}$$

Denote the correlation parameters by $\boldsymbol{\Theta} = (\alpha', \theta', \gamma')$, where $\alpha = (\alpha_1, \cdots, \alpha_t)'$, $\theta = (\theta_1, \cdots, \theta_q)'$, and $\gamma = (\gamma_{111}, \cdots \gamma_{qm_qk_q})'$. We can estimate these parameters from data and obtain the ordinary kriging predictor. This will be explained with an example in the next section.

3.5 Hard Turning Experiment

The objective of our experiment is to optimize a hard turning process with respect to cutting forces. Hard turning is a metal cutting process that produces machined parts out of hard materials with good dimensional accuracy, surface finish, and surface integrity. Minimizing cutting forces will help reduce the power requirements, elastic distortion of the workpiece, and tool wear; thus reducing the manufacturing cost and improving the quality of the machined part.

Nine factors are selected for experimentation, which include one branching factor, two nested factors, and six shared factors. The factors and their ranges are shown in Table 6. A 30-run orthogonal-maximin BLHD is generated by using the modified simulated annealing algorithm proposed by Joseph and Hung (2008). The optimal design matrix is given in Table 12. The branching factor (cutting edge shape, z_1), is labeled "1" and "2" to stand for chamfer and hone, respectively. Two nested factors (v_1 and v_2) are nested within the cutting edge shape. Recall that, if the cutting edge is chamfer, v_1 stands for chamfer angle and v_2 stands for chamfer land length, otherwise there is no factor.

The experiments are performed using a highly sophisticated finite element based machining simulation software *AdvantEdge*. This software models the underlying physics of metal cutting as a thermo-mechanical plastic deformation process and captures various material and geometric nonlinearities of the process. The theoretical basis of the simulation model can be found in Marusich and Ortiz (1995). The simulations are computationally intensive and require hours of running time for producing a single output (about 12 to 24 hours). The simulation outputs are deterministic and incorporate all the factors in Table 6. Various responses are produced by the software such as temperature, residual stresses, and forces. A finite element mesh and temperature distribution is shown in Figure 21. In this article, we chose to analyze only the resultant cutting force (y). The data are given in Table 12.

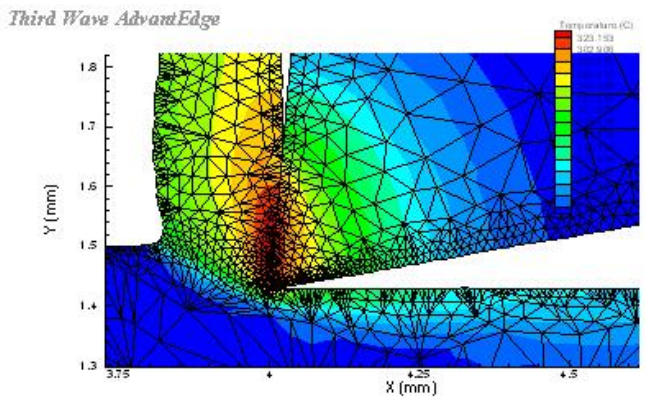

Figure 21: Finite element mesh and temperature distribution

This is an example where the number of nested factors are not the same for different levels of the branching factor. For simplicity of notations, we had assumed them to be the same in Section 3. Therefore, we should explain, how the criteria can be modified to deal with unequal number of nested factors. This is easy to do. Consider first the maximin criterion ϕ_λ. In (3), use $m_1 = 2$ for the first 15 runs and use $m_1 = 0$ for the last 15 runs. Now consider ρ^2 in (6). The pairwise correlations involving a nested factor should be calculated using the first 15 runs. Moreover, because there are no factors when branching factor level is "2", we do not need to consider the branching-by-nested interaction.

Since the cutting forces are positive, we first apply a log transformation before fitting the ordinary kriging model. We also normalize all of the factor settings in Table 12 into $[-1, 1]$. The 30 design points after normalization are denoted by $\{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_{30}\}$, where for all $1 \le j \le 30$, $\boldsymbol{w}_j = (x_{j1}, \cdots, x_{j6}, z_{j1}, v_{j1}^{z_{j1}}, v_{j2}^{z_{j1}})'$, $z_{j1} = -1$ represents the chamfer edge and $z_{j1} = 1$ represents the hone edge. The parameters in the kriging model can be estimated as (Santner, William, and Notz, 2003)

$$\begin{aligned}
\hat{\boldsymbol{\Theta}} &= arg\min_{\boldsymbol{\Theta}} N\log\hat{\sigma^2} + \log|\boldsymbol{\Psi}|, \\
\hat{\mu} &= (\mathbf{1}'\boldsymbol{\Psi}^{-1}\mathbf{1})^{-1}\mathbf{1}'\boldsymbol{\Psi}^{-1}\mathbf{y}, \\
\hat{\sigma}^2 &= \frac{1}{N}(\mathbf{y} - \hat{\mu}\mathbf{1})'\boldsymbol{\Psi}^{-1}(\mathbf{y} - \hat{\mu}\mathbf{1}),
\end{aligned}$$

Table 12: Orthogonal-maximin BLHD and data for the hard turning experiment

Run	z_1	v_1	v_2	x_1	x_2	x_3	x_4	x_5	x_6	y
1	1	1	6	15	23	7	9	18	10	162.1
2	1	2	11	25	3	25	14	25	19	284.9
3	1	3	3	4	20	18	18	5	26	160.3
4	1	4	14	9	6	6	27	7	17	121.1
5	1	5	8	16	8	21	2	2	1	104.6
6	1	6	1	17	10	5	25	19	25	241.9
7	1	7	12	29	26	15	5	14	12	195.4
8	1	8	5	26	16	30	22	15	6	159.5
9	1	9	15	7	13	26	7	11	27	241.6
10	1	10	10	1	29	20	23	6	5	88.33
11	1	11	2	20	21	27	10	20	29	320.4
12	1	12	7	8	11	14	4	29	21	218.8
13	1	13	13	22	9	1	24	27	9	193.5
14	1	14	4	10	2	24	28	13	13	198.6
15	1	15	9	28	25	13	17	3	28	155.1
16	2			19	5	9	1	8	20	164.4
17	2			14	28	17	6	21	24	323.6
18	2			6	17	4	16	12	4	109.1
19	2			11	1	12	15	4	8	115.4
20	2			27	22	8	30	24	16	254.8
21	2			21	14	23	19	10	22	217.0
22	2			23	18	22	12	28	3	243.7
23	2			3	27	3	3	26	14	131.5
24	2			13	15	19	29	16	30	258.7
25	2			24	12	2	11	1	18	109.3
26	2			18	24	28	8	17	2	174.8
27	2			12	30	11	26	9	11	157.0
28	2			2	4	16	13	30	15	133.1
29	2			30	7	10	20	23	7	210.1
30	2			5	19	29	21	22	23	273.3

where $\mathbf{1}$ is a vector of 1's having length 30, $\mathbf{y} = (y_1, \cdots, y_{30})'$, and $\mathbf{\Psi}$ is a 30×30 matrix whose njth element is

$$\exp\Big\{ -\textstyle\sum_{i=1}^{6} \hat{\alpha}_i (x_{ni} - x_{ji})^2 - \hat{\theta}_1 \mathbf{I}_{[z_{n1} \neq z_{j1}]} \\ -\hat{\gamma}_{111}(v_{n1}^{z_{n1}} - v_{j1}^{z_{j1}}) \mathbf{I}_{[z_{n1}=z_{j1}=-1]} - \hat{\gamma}_{121}(v_{n2}^{z_{n1}} - v_{j2}^{z_{j1}}) \mathbf{I}_{[z_{n1}=z_{j1}=-1]} \Big\}.$$

We obtain

$$\hat{\alpha} = (0.09, 0.01, 0.03, 0.01, 0.94, 1.08)',\ \hat{\theta} = \hat{\theta}_1 = 0.13,\ \hat{\gamma} = (\hat{\gamma}_{111}, \hat{\gamma}_{121})' = (0.14, 0.01)', \hat{\mu} = 5.1.$$

Note that we do not need to estimate γ_{112} and γ_{122} in this example, because there is no factor nested within hone edge. Thus, ordinary kriging predictor is given by (see, e.g., Joseph 2006)

$$\hat{y}(\boldsymbol{w}) = 5.1 + \hat{\boldsymbol{\psi}}(\boldsymbol{w})' \hat{\mathbf{\Psi}}^{-1} (\mathbf{y} - \mathbf{5.11}), \tag{12}$$

where $\boldsymbol{w} = (x_1, \cdots, x_6, z_1, v_1^{z_1}, v_2^{z_1})' \in [-1, 1]^9$ and $\hat{\boldsymbol{\psi}}(\boldsymbol{w})$ is a vector of length 30 with the jth element

$$\exp\Big\{ -\sum_{i=1}^{6} \hat{\alpha}_i (x_i - x_{ji})^2 - \hat{\theta}_1 \mathbf{I}_{[z_1 \neq z_{j1}]} - \hat{\gamma}_{111}(v_1^{z_1} - v_{j1}^{z_{j1}}) \mathbf{I}_{[z_1=z_{j1}=-1]} - \hat{\gamma}_{121}(v_2^{z_1} - v_{j2}^{z_{j1}}) \mathbf{I}_{[z_1=z_{j1}=-1]} \Big\}.$$

Based on equation (5) in Chapter two, the blind kriging predictor can be written as

$$\hat{y}(\boldsymbol{w}) = 5.1 + 0.2x_{5l} + 0.2x_{6l} - 0.12x_{1l}x_{6l} + \hat{\boldsymbol{\psi}}(\boldsymbol{w})' \hat{\mathbf{\Psi}}^{-1} (\mathbf{y} - \mathbf{V}_3 \hat{\mu}_m), \tag{13}$$

where $\hat{\alpha} = (0.73, 0.01, 0.62, 0.01, 0.12, 1.25)'$, $\hat{\theta} = \hat{\theta}_1 = 0.01$, $\hat{\gamma} = (\hat{\gamma}_{111}, \hat{\gamma}_{121})' = (0.01, 0.04)'$.

To understand the effects of the factors, we apply the sensitivity analysis technique on the ordinary kriging predictor (see Welch et al. 1992). The main effects plot is shown in Figure 22 (a). We can see that the cutting edge radius (x_1), feed (x_5), depth of cut (x_6), and chamfer angle (v_1) have significant effects on the cutting force. We also found a significant interaction between cutting edge radius and depth of cut (Figure 22 (b)). The depth of cut has a positive effect on the cutting forces, but

this effect is more significant when the cutting edge radius is smaller. This can be explained physically as follows. For a small cutting edge radius, an increase in depth of cut produces an increase in material deformation through shear and consequently its effect on the force is more significant. For larger cutting edge radius values, the contribution of ploughing of material around the cutting edge to the cutting force is more pronounced and consequently an increase in depth of cut does not produce as significant a change in the cutting force.

The optimal setting of the factors can be found by minimizing the ordinary kriging predictor in (12). We obtain $(x_1, x_2, x_3, x_4, x_5, x_6, z_1, v_1, v_2) = (-1.00, -0.76, 0.69, 0.70, -0.66, -0.70, -1, 0.05, 0.16)$, which is very close to the optimal setting found by minimizing the blind kriging predictor in (13). In their original scales, the optimal setting for the shared factors is $(x_1, x_2, x_3, x_4, x_5, x_6) = (5, -13.80, 1.41, 222, 0.067, 0.123)$ and the optimal cutting edge geometry is chamfer with angle 18.74 degree and length 128.13 microns. The resultant cutting force predicted under this setting is 81 N, which is much smaller than the observed forces in the experiment. We also performed a new experiment at the optimal setting and obtained the resultant force as 79 N. This confirms the validity of the optimal setting obtained from our model.

3.6 Conclusions

Design and analysis of experiments with branching and nested factors have surprisingly received scant attention in the literature. One possible reason for this could be that the experiments can be performed in two stages, i.e., in the first stage perform an experiment with the branching factors and shared factors. Because there are no nested factors, this experiment can be designed using the existing methods. Now the investigator can analyze the data and find out the best level of the branching factor. Then, a second stage of the experiments can be performed using only the nested factors under the optimum level of the branching factor. The design for this

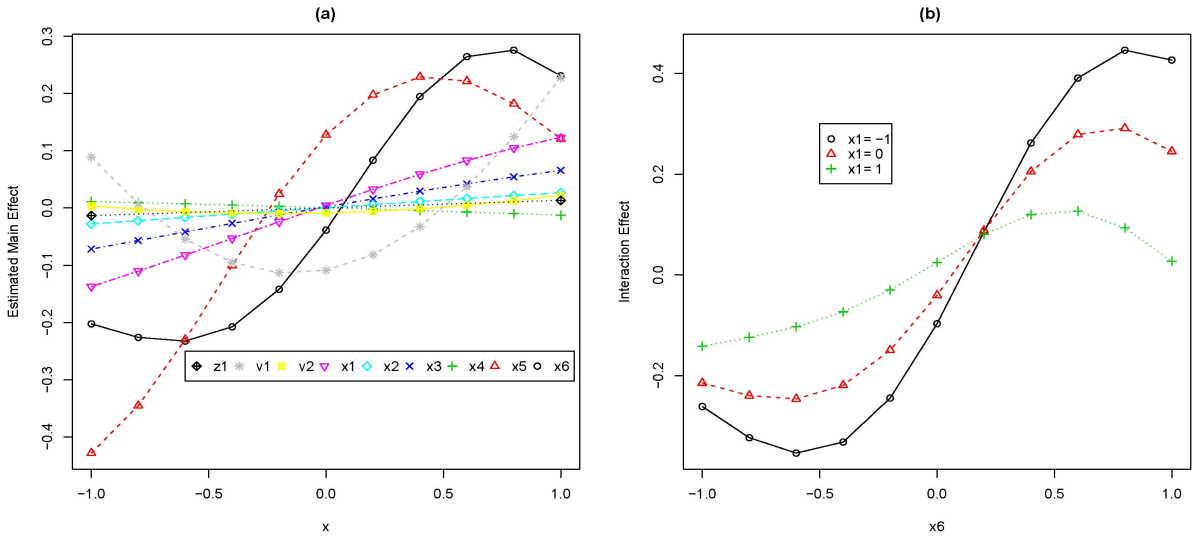

Figure 22: (a) Main effects plot and (b) interaction between x_1 and x_6.

experiment can also be easily obtained using the existing methods. Although this two-stage approach is quite intuitive, the final results may not be optimal. This is because, a different level of the branching factor may be the true optimum, but could not be identified in the first stage of the experiment because the nested factors under that branching level was not set at their optimal levels. This problem can be avoided using branching designs. It allows us to find the optimal settings of the branching factors, nested factors, and shared factors simultaneously.

Taguchi (1987) and Phadke (1989) have reported several case studies on experiments using branching designs, but their approach is not general enough to apply to more complex experiments such as a computer experiment. The optimality properties of their approach using orthogonal arrays are also not known. In this work, we have proposed branching Latin hypercube designs that is suitable for a computer experiment when it involves branching and nested factors. The optimal choice of such designs are also discussed. The approach was successfully applied to the optimization of a machining process.

Although the primary focus was on Latin hypercube designs, some issues regarding the use of orthogonal arrays and its applications in physical experiments are also important. Research on these topics is currently ongoing and will be reported elsewhere.

CHAPTER IV

BINARY TIME SERIES MODELING WITH APPLICATION TO ADHESION FREQUENCY EXPERIMENTS

[1]

4.1 Introduction

This research is motivated by the statistical analysis of time series data from biomechanical experiments that study protein, DNA, and RNA at the level of single molecules (Mehta et al., 1999). Single molecule mechanics experiments employ ultrasensitive force techniques to characterize mechanically a single pair of molecules that physically links the force sensor to a sample surface. Figure 1 illustrates a simple experiment - the micropipette adhesion frequency assay (Chesla et al., 1998). Here, a human red blood cell (Figure 23, A-C, left) pressurized by micropipette suction is used as a force transducer to test interactions between molecules presented on the red cell membrane and the counter molecules on the surface of another cell (Figure 23, A-C, right, only partly shown). The two cells are put together for a pre-determined duration (Figure 23B), then retracted away. The simplest measurement is whether a controlled contact results in adhesion. If adhesion is resulted, retraction will stretch the red cell (Figure 23C). If no adhesion is resulted, the red cell will not be stretched (Figure 23A). When adhesion does occur, additional quantities can be measured using the force transducer (Figure 23D).

To ensure adhesion to be mediated by a single molecular bond, the experimental

[1]The paper based on this chapter will appear in *Journal of American Statistical Association*.

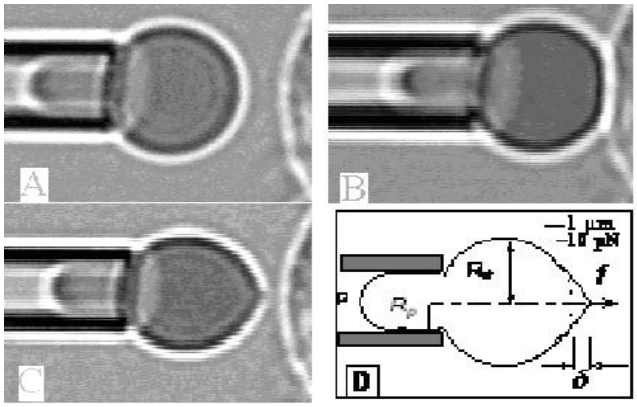

Figure 23: Photomicrographs (A-C) and schematics of micropipette adhesion frequency assay

condition is designed such that adhesion is infrequent (Zhu et al., 2002). As such, in any particular test both positive (i.e., adhesion, scored 1) as well as negative (i.e., no adhesion, scored 0) outcomes are possible and random. Due to the inherent stochastic nature of single molecular interactions, such analysis would require a large number of measurements to obtain their statistical properties. For example, the probability of adhesion can be estimated from the frequency of occurrence of adhesion in a large number of contacts (Chesla et al., 1998). The probability distribution of single bond lifetimes can be estimated from the histogram of a large number of lifetime measurements (Marshall et al., 2003). Experimentally, these are obtained by sequentially repeating the measurements many times.

A crucial assumption that allows measurements from repeated tests to be used for probability calculation is that all measurements are identical yet independent from each other, in other words, the test sequence consists of independent and identically distributed random variables. However, this may or may not be valid depending on the particular biological system in question. Recently, Zarnitsyna et al. (2007) demonstrated that this assumption is not valid in some biological systems. Specifically, it is shown that the occurrence of adhesion in the immediate past test can either

increase or decrease the likelihood for the next test to result in an adhesion. A simple analysis has been developed to determine whether the independent assumption is valid, and if not, to measure the amount of change in the probability of adhesion in the next test due to the occurrence of adhesion in the immediate past test (Zarnitsyna et al., 2007).

In this article, we extend the simply analysis to a more sophisticated binary time series model. Numerous methods for binary time series analysis are available in the literature (Zeger and Qaqish, 1988; Li, 1994; Slud and Kedem, 1994; Benjamin et al., 2003). Most of these methods are developed for a single series of observations. Extensions to multiple binary time series modeling and related inferences have not been systematically studied. Both Li (1994) and Kedem and Fokianos (2002, p. 84) pointed out the importance of extensions to cases where a series is collected for each individual. This is different from classical time series analysis in that the binary time series are observed on different replicates of the experimental units. Correlation among the repeated observations may arise not only from memory effects but also from shared unobserved variables. Therefore, more general models are required to incorporate the correlations among repeated observations. Another important issue is model diagnostic. In distinction to Pearson's χ^2 test which works under the independence assumption, new test statistics and their theoretical properties need to be developed.

The remainder of this article is organized as follows. Some preliminary analysis results for an adhesion frequency experiment are presented in Section 2. In Section 3, a class of multiple binary time series models is proposed. A goodness-of-fit test for model assumptions and their asymptotic properties are derived in Section 4 and its finite-sample performance is examined via a simulation study. In Section 5, the proposed model and inferences are applied to the same experiment and the results are compared with those in Section 2. Summary and concluding remarks are given

in Section 6.

4.2 Preliminary Analysis of an Adhesion Frequency Experiment

In the micropipette adhesion frequency assay, adhesion between the two cells are staged by placing them onto controlled contact with given contact time and area via a computer-driven micromanipulation to ensure each contact was as close to identical to any other contacts as possible (Figure 23). Average number of bonds (ANB) is a transformation of the contact time. It can be calculated based on a chemical equation (Chesla et al., 1998)

$$ANB = A_c m_r^{\nu_r} m_1^{\nu_1} K_a^0 [1 - exp(-k_r^0 CT)], \tag{1}$$

where CT is the corresponding contact time, A_c, $m_r^{\nu_r}$, $m_1^{\nu_1}$, K_a^0, and k_r^0 are biological constants representing the densities of the interacting molecules and their binding kinetic rates. For each average number of bonds, several replicates of cell pairs are tested in the experiment. For each pair of cells, adhesion test cycle (i.e. contact and retraction) was repeated 50 times. Test scores (denoted by y) are recorded in binary form (i.e., $y = 0$ or 1), which results in multiple binary time series of the type exemplified in Table 13.

Under independent Bernoulli trial assumption (Chesla et al., 1998), the average adhesion probability (P_{ANB}) can simply be estimated by the adhesion frequency calculated as

$$P_{ANB} = \frac{\text{number of adhesions}}{\text{number of test cycles}}. \tag{2}$$

Figure 24 shows an example of the relationship between adhesion probability and average number of bonds (Zhang and Zhu, unpublished data). In this micropipette experiment, the adhesion test was conducted with seven different average number of bonds (0.085, 0.17, 0.255, 0.34, 0.51, 0.68, and 1.36). For the first two average number

Table 13: Example of adhesion frequency experiment data

Average Number of Bonds (ANB)	50 Repeated Adhesion Tests
0.085	$01010011011101010000\cdots$
0.085	$00010000100010100110\cdots$
0.085	$10000100100010000101\cdots$
$\vdots$	$\vdots$
0.170	$10110000110001111101\cdots$
0.170	$10001000000000000011\cdots$
0.170	$00010000100101110011\cdots$
$\vdots$	$\vdots$

of bonds (0.085 and 0.17), each has six pairs of replicates; for the remaining, there are five pairs each. Each point in Figure 24 represents the P_{ANB} value for one pair of cells, and is calculated from equation (1). The solid line represents the average over all the replicates under the same average number of bonds.

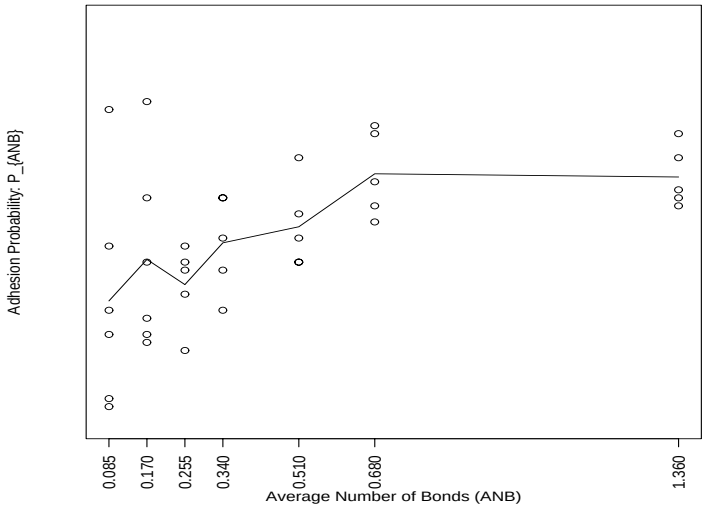

Figure 24: Adhesion probability (P_{ANB}) varies with the average number of bonds (ANB)

To understand the relationship between P_{ANB} and ANB, the existing method (Chesla et al., 1998) is based on the assumption that the binary time series data (e.g., Table 13) form Bernoulli sequences. However, for each pair of cells, the adhesion

test cycles are observed repeatedly. The independence assumption may not hold as recently demonstrated (Zarnitsyna et al., 2007). Therefore, it is necessary to check the adequacy of the distributional assumption before applying the method. One graphical technique to assess this assumption is the probability plot. If the data are collected from independent Bernoulli trials, the number of trials (i.e., tests) needed to get one success (i.e. 1) will follow a geometric distribution with probability p, where $p = \text{Prob}(y = 1)$. Figure 25 includes four different average number of bonds (0.085, 0.17, 0.255, 0.34). For each average number of bonds, the numbers of tests needed to get one success are calculated over all replicates. Then, its empirical cumulative distribution are plotted against the geometric distribution, where the parameter p is estimated by (2) at each average number of bonds. In the top two panels, significant deviations from the straight line indicate violation of the independent Bernoulli assumption, and the departure increases as the average number of bonds decreases. Similar conclusions were first observed by using a different analysis in Zarnitsyna et al. (2007) which also motivated our present work.

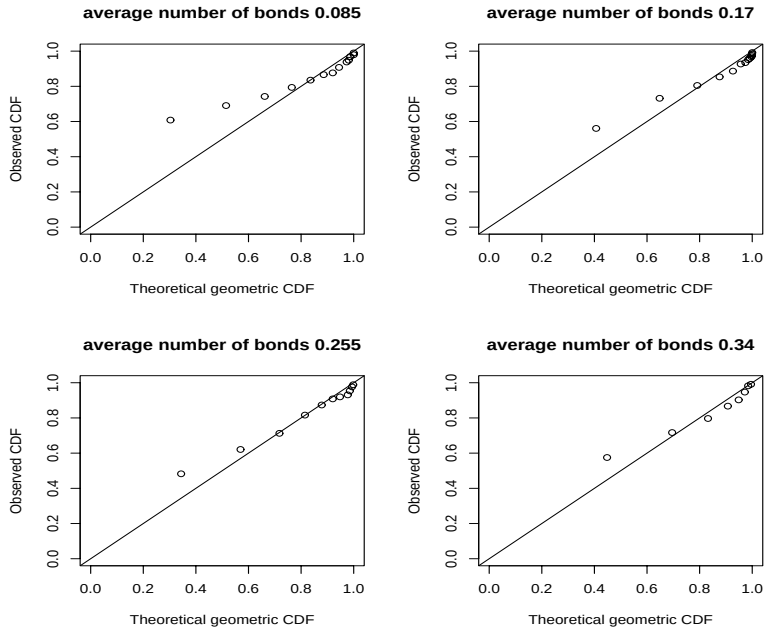

Figure 25: Probability plot

To gain further insight on the violation of the independent Bernoulli assumption, additional graphical plots are used here to better understand the dependence among repeated binary observations. The idea is to compare the conditional adhesion probability given the previous test results. Define P(1|1) to be the conditional adhesion probability given adhesion in the previous test, and P(1|0) the conditional probability given no adhesion in the previous test. If the test results are independent, P(1|1) should be equal to P(1|0) and both can be estimated by P_{ANB} in (2). In Figure 26, for each average number of bonds, the green points represent the conditional probability P(1|1) calculated for each replicate. The green line stands for P(1|1) calculated over all replicates under the same average number of bonds. Similarly, the red points and red line are those for the conditional probability P(1|0). For comparison, the black lines shows the adhesion probability P_{ANB} calculated by (2) at each average number of bonds. As the green line and points are much higher than the red ones, the adhesion probability is higher if adhesion occurs in the previous test. This lends strong evidence for *memory effect* on repeated tests. A more in-depth biological discussion can be found in Zarnitsyna et al. (2007), where the *memory effect* was first observed by using a different analysis. From Figure 26, one can visually infer about the existence of serial correlation and interactions. The heterogeneity among subjects is also transparent from Figure 26. To describe and quantify significant effects on the adhesion probability, one should consider the use of a new binary time series model which incorporates the various effects suggested by the plots.

4.3 Modeling and Estimation

4.3.1 Modeling

In this section, a new binary time series model will be proposed. First, we need to review some existing models.

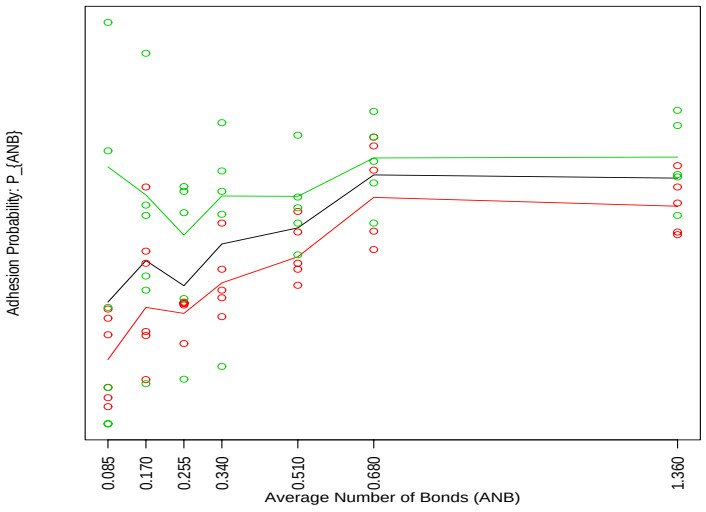

Figure 26: Memory effects in micropipette experiments

4.3.1.1 Random effects models

Random effects models are most useful in longitudinal data analysis when correlation arises from some unobservable variables shared among repeated observations. Consider a binary realization $\{y_{ij}\}$ taking values 0 or 1 for subject i at jth observation. For given subject-specific coefficients β_i, assuming the repeated observations for each individual are independent, the random effects model takes the form

$$log\frac{Pr(y_{ij}=1 \mid \beta_i)}{1-Pr(y_{ij}=1 \mid \beta_i)} = \beta_0 + \beta_i + x'_{ij}\alpha, \tag{3}$$

where the vector x_{ij} denotes the covariates associated with the fixed effects α, and the random effects β_i's are mutually independent with a common underlying multivariate distribution. This model is used to represent the natural heterogeneity across individuals in the regression coefficient. More discussion about this model can be found in Diggle et al. (2002).

4.3.1.2 Binary time series models

Non-Gaussian time series modeling techniques has been extensively discussed in the literature. Benjamin et al. (2003) proposed a generalized autoregressive moving

average (GARMA) model. Applying this GARMA model with logistic link, a binary time series $\{y_t\}$ can be fitted as

$$logit(\mu_t) = x_t'\alpha + \sum_{r=1}^{R} \varphi_r \mathcal{A}(y_{t-r}) + \sum_{q=1}^{Q} \zeta_q \mathcal{M}(y_{t-q}, \mu_{t-q}), \tag{4}$$

where x_t are covariates at time t, $\mu_t = E(y_t \mid H_t)$ is the conditional mean given the previous information $H_t = \{x_t, \cdots, x_1, y_{t-1}, \cdots, y_1, \mu_{t-1}, \cdots, \mu_1\}$. $\mathcal{A}$ and $\mathcal{M}$ are functions representing the autoregressive (AR) and moving average (MA) terms with corresponding order R and Q. These two functions together are denoted by ARMA(R,Q). φ_r's and ζ_q's are the AR and MA parameters. For binary time series, a reasonable choice for $\mathcal{A}$ and $\mathcal{M}$ can be respectively y_t and residuals such as $y_t - \mu_t$.

Model (4) includes many well-known models as special cases. One important submodel is the Zeger-Qaqish model with logistic link

$$logit(\mu_t) = x_t'\alpha + \sum_{r=1}^{R} \varphi_r y_{t-r}, \tag{5}$$

and the moving average form for this model (Li, 1994)

$$logit(\mu_t) = x_t'\alpha + \sum_{q=1}^{Q} \zeta_q \bigg(y_{t-q} - \mu_{t-q} \bigg). \tag{6}$$

Inference and asymptotic properties for the autoregressive logistic regression models are discussed via conditional likelihood (Kaufmann, 1987) and partial likelihood (Slud and Kedem, 1994; Kedem and Fokianos, 2002).

We propose a *binary time series mixed model* (BTSM). It is a multiple logistic time series model with random effects that takes into account the heterogeneity among experimental units. Consider a binary time series realization $\{y_{it}\}$ taking values 0 or 1 for subject i at time t, where $i = 1, \cdots, m$, $t = 1, \cdots, n$, and $mn = N$. Suppose the experimental units are sampled from a population. It is reasonable to assume that the random effects β_i's are independent from a normal distribution with mean b and variance σ_b^2. For the vector $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_m)'$, its distribution can be written as

$\mathcal{N}(\mathbf{b}, \mathbf{\Sigma})$, where $\mathbf{b}$ is a column of b's having length m, $\mathbf{\Sigma} = \sigma_b^2 \mathbf{I}_m$, and $\mathbf{I}_m$ is the $m \times m$ identity matrix. The vector $x_{it} = \{x_{it,1}, \cdots, x_{it,p}\}'$ denotes the covariates associated with the p-dimensional fixed effects $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_p)'$, and $z_{it} = \{z_{it,1}, \cdots, z_{it,m}\}'$ the design matrix for the random effects $\boldsymbol{\beta}$ such that $z_{it}'\boldsymbol{\beta} = \beta_i$, that is $z_{it,i} = 1$ and $z_{it,j} = 0$ for all $j \neq i$.

Denote the conditional mean $\mu_{it} = \mathrm{E}(y_{it} \mid H_{it})$. Given the previous information $H_{it} = \{x_{it}, x_{it-1}, x_{it-2}, \cdots, y_{it-1}, y_{it-2}, \cdots, \mu_{it-1}, \mu_{it-2}, \cdots\}$ and random effects, y_{it} are conditionally independent with mean $\mathrm{E}(y_{it}|\boldsymbol{\beta}, H_{it}) = \mu_{it}^{\boldsymbol{\beta}}$. By logistic link function, the conditional mean $\mu_{it}^{\boldsymbol{\beta}}$ is related to the linear predictor $\eta_{it}^{\boldsymbol{\beta}}$ by

$$logit(\mu_{it}^{\boldsymbol{\beta}}) = \eta_{it}^{\boldsymbol{\beta}} = z_{it}'\boldsymbol{\beta} + x_{it}'\boldsymbol{\alpha} + \sum_{l=1}^{L} \gamma_i x_{it-l} y_{it-l} + \sum_{r=1}^{R} \varphi_r y_{it-r} + \sum_{q=1}^{Q} \zeta_q (y_{it-q} - \mu_{it-q}). \quad (7)$$

This model is called a BTSM model. The random effects $\boldsymbol{\beta}$ are used to represent a variety of situations, including subject heterogeneity, unobserved covariates, and other forms of overdispersion. Here the heterogeneity is modelled directly through subject-specific parameter. If random intercept along may not sufficiently capture the variation exhibited in the data, this model can be easily extended to a general form by incorporating more complicated random effects. Given β_i, the y_{it}'s are correlated because y_{it-l} explicitly influence y_{it}. This correlation can be explain by the AR and MA components in (7). The MA process which involves $\mu_{i,t-q}$ makes the model more complicated. In this formulation, the interaction terms $(x_{it-1}y_{it-1}, \cdots, x_{it-L}y_{it-L})$ between covariates and past outcomes provide flexibility in adjusting the time series structure with respect to different covariates settings.

The proposed BTSM model is general and includes the models discussed heretofore. The random effects model in (3) is a submodel of BTSM under the assumption that the repeated measurements for each unit are independent, and the correlation among repeated observations arises only from some unobserved variables. With the

logistic link function, the GARMA model in (4) is a special case of BTSM if no random effect is included. That is, based on the population average, it models the time series structure without considering the heterogeneities among the units.

More than a simple extension of existing models, the BTSM model poses some challenging tasks. By considering the hidden variables shared among units, it incorporates random effects in logistic time series regression. This makes the estimation and inference more complicated and different from standard binary time series analysis. Another important issue is the goodness-of-fit test for model diagnostic. There are related work for linear mixed models in the literature (Jiang, 2001a,b). There is, however, no existing method for testing the distributional assumption in binary time series models with random effects. Furthermore, the asymptotic χ^2 distribution cannot be applied to the new test statistics because of its independence assumption. Instead, a martingale central limit theorem will be used in the next section to derive the asymptotic properties.

4.3.2 Estimation by Partial Likelihood

Model fitting procedure herein is based on partial likelihood (PL). PL was introduced by Cox (1972, 1975). More formal definition and theoretical justification can be found in Wong (1986) and Slud (1992). The advantage for PL is that it enables very flexible conditional inferences for all practical purposes, especially when time-dependent covariates are involved, i.e., H_{it} may not include the information for covariates at time t. Fokianos and Kedem (2004) have discussions on using PL in time series which follow generalized linear models. For the BTSM model, the presence of random effects causes some integration difficulty, which makes the estimation different from standard methods in time series analysis. In this section, an approximation procedure will be proposed to tackle this problem.

Denote the observation vector by $\mathbf{y} = (y_1, \cdots, y_m)'$, where the observations for subject i are $y_i = (y_{i1}, \cdots, y_{in})'$, and $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_O)'$, $\boldsymbol{\varphi} = (\varphi_1, \cdots, \varphi_R)'$, $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_Q)'$. Assume

$$\omega = (\boldsymbol{\alpha}', \boldsymbol{\gamma}', \boldsymbol{\varphi}', \boldsymbol{\zeta}')'$$

are s-dimensional fixed effects, and $\mathbf{X}$ the corresponding matrix with rows

$$X_{it}' = (x_{it}', x_{i\,t-1}y_{i\,t-1}, \cdots, x_{i\,t-O}y_{i\,t-O}, y_{i\,t-1}, \cdots, y_{i\,t-R}, (y_{i,t-1}-\mu_{it-1}), \cdots, (y_{i,t-Q}-\mu_{it-Q})).$$

Similarly, with rows z_{it}', the design matrices for the random effects are denoted by $\mathbf{Z}$. Given the previous information H_{it}, the corresponding partial likelihood for fixed effects is

$$PL(\omega|\boldsymbol{\beta}) = \prod_{i=1}^{m}\prod_{t=1}^{n} pl_{\omega}(y_{it}|\boldsymbol{\beta}, H_{it}) = \prod_{i=1}^{m}\prod_{t=1}^{n} [\pi_{it}(\omega|\boldsymbol{\beta})]^{y_{it}}[1-\pi_{it}(\omega|\boldsymbol{\beta})]^{1-y_{it}},$$

where $\pi_{it}(\omega|\boldsymbol{\beta}) = P_{\omega|\boldsymbol{\beta}}(y_{it} = 1 \mid H_{it}) = \mu_{it}^{\boldsymbol{\beta}}$.

The integrated quasi-partial likelihood function used to estimate (ω, σ_b^2) is defined by

$$\mid \boldsymbol{\Sigma} \mid^{-1/2} \int exp\left[logPL(\omega|\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right] d\boldsymbol{\beta}$$

$$= \mid \boldsymbol{\Sigma} \mid^{-1/2} \int exp\left[\sum_{i=1}^{m}\sum_{t=1}^{n}(y_{it}log\frac{\pi_{it}(\omega|\boldsymbol{\beta})}{1-\pi_{it}(\omega|\boldsymbol{\beta})} + log(1-\pi_{it}(\omega|\boldsymbol{\beta}))) - \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right] d\boldsymbol{\beta}.$$

Because of the difficulty in implementing the full partial likelihood, we use penalized quasi-partial likelihood (PQPL) as an approximation. PQPL is an extension of penalized quasi-likelihood (Breslow and Clayton, 1993), which has generally been used to circumvent the same integration difficulty for a generalized linear mixed model. The idea is to apply Laplace's method for integral approximation (Barndorff-Nielsen and Cox, 1989, Sec. 3.3; Tierney and Kadane, 1986). Hence, the integrated quasi-partial log-likelihood can be approximated by

$$ql(\omega, \sigma_b) \approx -\frac{1}{2}log|\mathbf{I}_m + \mathbf{Z}^t\mathbf{W}\mathbf{Z}\boldsymbol{\Sigma}| + \sum_{i=1}^{m}\sum_{t=1}^{n}(y_{it}log\frac{\pi_{it}(\omega|\tilde{\boldsymbol{\beta}})}{1-\pi_{it}(\omega|\tilde{\boldsymbol{\beta}})} + log(1-\pi_t(\omega|\tilde{\boldsymbol{\beta}}))) - \frac{1}{2}\tilde{\boldsymbol{\beta}}'\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{\beta}}, \tag{8}$$

where $\mathbf{W}$ is the $N \times N$ diagonal matrix with diagonal terms $w_{it} = \pi_{it}(\omega|\tilde{\boldsymbol{\beta}})(1 - \pi_{it}(\omega|\tilde{\boldsymbol{\beta}}))$ and $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\omega, \sigma_b)$ is the solution of $\sum_{i=1}^{m}\sum_{t=1}^{n}(y_{it} - \pi_{it}(\omega|\boldsymbol{\beta}))z_{it} - \frac{\boldsymbol{\beta}}{\sigma_b^2} = 0$, which maximizes the sum of the last two terms in (8). Following the assumption in Breslow and Clayton (1993) that the GLM iterative weights vary slowly as a function of mean, the first term in (8) can be ignored. So ω is chosen to maximize the second term. That is, $(\hat{\omega}, \hat{\boldsymbol{\beta}}) = (\hat{\omega}(\sigma_b), \hat{\boldsymbol{\beta}}(\sigma_b))$, where $\hat{\boldsymbol{\beta}}(\sigma_b) = \tilde{\boldsymbol{\beta}}(\hat{\omega}(\sigma_b))$ jointly maximize

$$\sum_{i=1}^{m}\sum_{t=1}^{n}(y_{it} log\frac{\pi_{it}(\omega, \sigma_b)}{1 - \pi_{it}(\omega, \sigma_b)} + log(1 - \pi_{it}(\omega, \sigma_b))) - \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}, \tag{9}$$

with $\pi_{it}(\omega, \sigma_b) = P_{\omega,\sigma_b}(y_{it} = 1 \mid H_{it})$.

Differentiation of (9) with respect to fixed effects ω and random effects $\boldsymbol{\beta}$ leads to the penalized quasi-partial score equations:

$$\sum_{i=1}^{m}\sum_{t=1}^{n} X_{it}(y_{it} - \pi_{it}(\omega, \sigma_b)) = 0, \tag{10}$$

$$\sum_{i=1}^{m}\sum_{t=1}^{n} z_{it}(y_{it} - \pi_{it}(\omega, \sigma_b)) = \boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}. \tag{11}$$

Given σ_b, the maximum quasi-partial likelihood estimator(MQPLE) of $(\hat{\omega}, \hat{\boldsymbol{\beta}})$ can be obtained by solving these two score equations. An important role in partial likelihood inference is played by the score process (10) and (11), which is a vector of martingales with respect to H_{it}. Hence, in Section 3.4 the study of asymptotic behavior of the MQPLE $\hat{\omega}$ will be based on central limit theorems for martingales. Questions regarding existence and uniqueness of the MQPLE are important, because the score equations (10) and (11) may have multiple roots, or they may have no roots at all. Similar questions for the traditional maximum likelihood estimators (MLE) have been addressed by a number of authors. Silvapulle (1981) and Albert and Anderson (1984) provide some necessary and sufficient conditions for the existence of MLE for binomial response models. Wedderburn (1976) and Kaufmann (1987) both give uniqueness conditions for the MLE. These results can also be applied to MQPLE and provide the essential conditions needed for existence and uniqueness of MQPLE.

Substitution of the maximized value of (9) from penalized quasi-partial likelihood into (8), and evaluating $\mathbf{W}$ at $(\hat{\omega}(\sigma_b), \hat{\boldsymbol{\beta}}(\sigma_b))$ generates an approximate profile quasi-likelihood function for the variance components. Using the same assumption in Breslow and Clayton (1993), we can approximate the profile quasi-likelihood function for inference on the variance component σ_b^2 by the working dependent variables $\mathbf{Y}$, iterated weights $\mathbf{W}$ and design matrices $\mathbf{X}$ and $\mathbf{Z}$. Define $\mathbf{V} = \mathbf{W}^{-1} + \mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}'$ and $\mathbf{Y}$ is a vector whose components are $Y_{it} = \eta_{it}^{\boldsymbol{\beta}} + \frac{(y_{it} - \mu_{it}^{\boldsymbol{\beta}})}{\mu_{it}^{\boldsymbol{\beta}}(1-\mu_{it}^{\boldsymbol{\beta}})}$. Up to an additive constant,

$$ql(\hat{\omega}(\sigma_b), \sigma_b) \approx -\frac{1}{2} log|\mathbf{V}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\hat{\omega})'\mathbf{V}^{-\mathbf{1}}(\mathbf{Y} - \mathbf{X}\hat{\omega}). \tag{12}$$

Restricted maximum likelihood (REML) (Patterson and Thompson, 1971) is used for estimation in (12) because it takes into account the loss in degrees of freedom resulting from estimating the fixed effects. Details can be found in Harville (1977) and Searle, et al. (1992). The REML version of (12) can be written as

$$ql(\hat{\omega}(\sigma_b), \sigma_b) \approx -\frac{1}{2} log|\mathbf{V}| - \frac{1}{2} log|\mathbf{X}'\mathbf{V}^{-\mathbf{1}}\mathbf{X}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\hat{\omega})'\mathbf{V}^{-\mathbf{1}}(\mathbf{Y} - \mathbf{X}\hat{\omega}). \tag{13}$$

Differentiating (13) with respect to σ_b^2 gives the estimating equation for the variance components:

$$-\frac{1}{2}\left[(\mathbf{Y} - \mathbf{X}\omega)'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_b^2}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\omega) - tr\left(P\frac{\partial \mathbf{V}}{\partial \sigma_b^2}\right)\right] = 0, \tag{14}$$

where $P = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$ (Harville, 1977).

Estimation of the fixed effects and variance components can be obtained by iteratively solving (10), (11) and (14). This estimation procedure is different from standard GLMM since the new model involves time series structure, i.e., the μ_{it-q} term depends on all the previous observations throughout the iterations. We may compute μ_{it-q} by setting initial μ_{it-q}'s to zero or to the sample mean of y_{it}. This should have negligible effect for a long enough iteration. Estimation can be carried out by simple modification in standard statistical software for GLMM such as SAS GLIMMIX package. Details about GLMM can be found in Breslow and Clayton

(1993). Questions regarding robust estimation and efficient algorithm have been addressed by a number of authors (McCulloch, 1997; Lin and Breslow, 1996; Pan, 2001).

4.3.3 Asymptotic Properties

Large sample properties for fixed effects and variance components in BTSM model are studied in this section. Considering a model which includes time-dependent covatiates, Fokianos and Kedem (2004) studied the asymptotic behavior of fixed effects ω in generalized linear time series models using partial likelihood inference. Theorem 1 is an extension of Fokianos and Kedem (2004) to multiple binary time series models with random effects. Based on the quasi-partial likelihood, Theorem 1 gives the consistency and asymptotic normality for the fixed effects estimators $\hat{\omega}$. With the help of the working dependent variables Y defined in Section 3, Theorem 2 gives the asymptotic properties for the REML estimator of σ_b^2 based on some asymptotic properties for linear mixed models with GLM iterative weights (Jiang, 1996). Assumptions and proofs are given in the appendix.

THEOREM 1. *Under assumptions A1 and A2, the maximum quasi-partial likelihood estimator (MQPLE) for the fixed effects $\hat{\omega}$ are consistent and asymptotically normal as $N \to \infty$:*

$$\sqrt{N}(\hat{\omega} - \omega) = \Lambda_N^{-1} \frac{1}{\sqrt{N}} S_n(\omega, \sigma_b) + o_p(1), \tag{15}$$

$$\sqrt{N}\Lambda_N^{1/2}(\hat{\omega} - \omega) \to_d \mathcal{N}(\mathbf{0}, I_s), \tag{16}$$

where the sample information matrix $\Lambda_N = \frac{1}{N}\sum_{t=1}^{n}\sum_{i=1}^{m} X_{it}X_{it}'\pi_{it}(\omega,\sigma_b)(1-\pi_{it}(\omega,\sigma_b))$, and $S_n(\omega,\sigma_b) = \sum_{t=1}^{n}\sum_{i=1}^{m} X_{it}(y_{it} - \pi_{it}(\omega,\sigma_b)) = \mathbf{X}'(\mathbf{y} - \pi(\omega,\sigma_b))$.

With the profile quasi-likelihood function (13), inference on variance component in model (7) can be formulated as an iterative procedure to estimate linear mixed model with the GLM iterative weight $\mathbf{W}^{-1}$ as

$$\mathbf{Y} = \mathbf{X}\omega + \mathbf{Z}\boldsymbol{\beta} + \epsilon, \tag{17}$$

where $\boldsymbol{\beta}$ comes from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $\epsilon = (\epsilon_1, \cdots, \epsilon_N)$ follows $\mathcal{N}(\mathbf{0}, \mathbf{W}^{-1})$, and the corresponding w_{it}'s are rewritten as $(w_1, \cdots, w_N)$. Recall that $\boldsymbol{\Sigma} = \sigma_b^2 \mathbf{I}_m$. Jiang (1996) developed rigorous asymptotic properties for REML estimates of variance components in linear mixed model (LMM) without the normality assumption on random effects and errors. Hence, Theorem 2 is a special case of Jiang (1996) for LMM with known unequal weights. Here we borrow some notation from Jiang (1996). Define

$$\mathbf{V}^* = A(A^t \mathbf{W}^{1/2} \mathbf{V} \mathbf{W}^{1/2} A)^{-1} A^t,$$

where A is any $N \times (N - s)$ matrix such that $\text{rank}(A) = N - s$ and $A^t \mathbf{W}^{1/2} \mathbf{X} = 0$,

$$g(\sigma_b) = (I_N, \sqrt{\sigma_b^2} \mathbf{W}^{1/2} \mathbf{Z})',$$

$$\varpi_l = \begin{cases} \epsilon_l \sqrt{w_l}\,, & 1 \le l \le N, \\ \frac{\beta_{l-N}}{\sqrt{\sigma_b^2}}\,, & N+1 \le l \le N+m, \end{cases}$$

$$\mathbf{V}_1 = (A' \mathbf{W}^{1/2} \mathbf{V} \mathbf{W}^{1/2} A)^{-1/2} A' \mathbf{W}^{1/2} \mathbf{Z}\mathbf{Z}' \mathbf{W}^{1/2} A (A' \mathbf{W}^{1/2} V \mathbf{W}^{1/2} A)^{-1/2},$$

$$V_1(\sigma_b) = g(\sigma_b) \mathbf{V}^* \mathbf{W}^{1/2} \mathbf{Z}\mathbf{Z}' \mathbf{W}^{1/2} \mathbf{V}^* g(\sigma_b)',$$

$$I^N = \frac{\text{tr}(\mathbf{V}_1 \mathbf{V}_1)}{m},\ K = \frac{1}{m} \sum_{l=1}^{N+m} (E\varpi_l^4 - 3) V_1(\sigma_b)_{ll}^2\,,$$

$$J = 2I^N + K.$$

THEOREM 2. *Under assumptions A3 and A4, as $N \to \infty$ and $m \to \infty$, the REML estimate for variance components is consistent and asymptotically normal with*

$$J^{-1/2} I^N \sqrt{m}(\hat{\sigma}_b^2 - \sigma_b^2) \to_d \mathcal{N}(0, 1). \tag{18}$$

4.4 Goodness-of-fit for Model Diagnostics

4.4.1 Goodness-of-fit Test

Pearson's χ^2 test is generally used to test if data follow some specific distribution. An important assumption for this test is the independence of the observations. How to

perform testing for model assumptions when the data come from a binary time series model? One approach is to classify the responses according to mutually exclusive events in terms of the previous output and then check the difference between observed and theoretical frequencies in each category. This can be written as follows.

Assume the binary data y_{it} comes from a binomial distribution with probability p depending on H_{it}. H_{it}, defined in Section 3.1, can be decomposed into several mutually exclusive events. Recall that $H_{it} = \{x_{it}, x_{it-1}, x_{it-2}, \cdots, y_{it-1}, y_{it-2}, \cdots, \mu_{it-1}, \mu_{it-2}, \cdots\}$. Suppose the decomposition is decided by n_1 covariates $(x_{it}, \cdots, x_{i\,t-n_1})$ decomposed into c_1 exclusive subsets, n_2 autoregression effects $(y_{i\,t-1}, \cdots, y_{i\,t-n_2})$ decomposed into c_2 exclusive subsets, and n_3 moving average effects $(\mu_{i\,t-1}, \cdots, \mu_{i\,t-n_3})$ decomposed into c_3 exclusive subsets, where $c_i \geq 1$, for $i = 1, 2, 3$. Therefore, there are K exclusive events denoted by $E_1, \ldots, E_K$ with

$$K = c_1 c_2 c_3, \tag{19}$$

and define

$$M_k \equiv \sum_{i=1}^{m} \sum_{t=1}^{n} y_{it} \mathbf{1}_{(H_{it} \in E_k)},$$

$$e_k(\omega, \sigma_b) \equiv \sum_{i=1}^{m} \sum_{t=1}^{n_i} \mathbf{1}_{(H_{it} \in E_k)} P_{\omega,\sigma_b}(y_{it} = 1 \mid H_{it}).$$

Similar to Pearson's χ^2-test, a test statistic can be defined by

$$\chi^* \equiv \sum_{k=1}^{K} \frac{(M_k - e_k(\omega, \sigma_b))^2}{\mathrm{E}(M_k)}. \tag{20}$$

If the parameters ω and σ are known, the asymptotic distribution for this test statistic is χ^2 with K degrees of freedom.

For the BTSM model, new construction and asymptotic properties of the goodness-of-fit test need to be rigorously established for two reasons. First, the probability $P_{\omega,\sigma_b}(y_{it} = 1 \mid H_{it})$ is not completely specified under the null hypothesis because it involves the unknown parameters (ω, σ_b). After replacing the unknown parameters in $e_j(\omega, \sigma_b)$ by the estimated values $(\hat{\omega}, \hat{\sigma_b})$, the χ^2 approximation may not be valid

(Chernoff and Lehmann, 1954; Jiang, 2001b). Second, because of the random effects and time series structures in the BTSM model, the observations are correlated. Accordingly, the asymptotic χ^2 result may not follow from the classic central limit theorem.

Jiang (2001b) derived the asymptotic distribution for goodness-of-fit test in linear mixed models (LMM) with continuous response to assess the adequacy of distributional assumptions. A new test statistic is constructed here based on binary observations and the corresponding time series model. Furthermore, the BTSM model includes time-dependent covariates. For this general formulation, inference is made based on the partial likelihood function. Asymptotic distribution of the new test statistic is derived by exploiting the martingale properties of the quasi-partial score process, which is different from Jiang (2001b).

Define a new goodness-of-fit test statistics for distributional assumptions in the BTSM model as

$$\hat{\chi}^2 = \frac{1}{N}\sum_{k=1}^{K}(M_k - e_k(\hat{\omega}, \hat{\sigma}_b))^2. \tag{21}$$

Unlike the Pearson's χ^2 test, the asymptotic distribution for this new statistic may not be χ^2. Hence, there is no need to have a normalizing constant in the test statistic to achieve χ^2 distribution. Instead, for simplicity, we choose a unified N as suggested in Jiang (2001b).

The asymptotic properties for the test statistic (21) are given in Theorem 3. Proofs are given in the appendix. The following notation is used in the theorem. Define

$$\theta = (\omega', \sigma_b),\ D = \left[\frac{1}{N}\sum_i\sum_t \mathbf{1}_{(H_{it}\in E_k)}\frac{\partial}{\partial\omega'}\pi_{it}(\theta)\right]_{1\le k\le K}\Lambda_N^{-1},$$

$$G_{it} = [\mathbf{1}_{(H_{it}\in E_k)} - DX_{it}]_{1\le k\le K},$$

$$C = \frac{1}{\sqrt{m}}\mathbf{W}^{1/2}\mathbf{V}^*\mathbf{W}^{1/2}\mathbf{Z}\mathbf{Z}'\mathbf{W}^{1/2}\mathbf{V}^*\mathbf{W}^{1/2},$$

$$\Phi = \frac{(I^N)^{-1}}{\sqrt{m}}\left[\sum_i\sum_t(\mathbf{1}_{(H_{it}\in E_k)}\frac{\partial}{\partial\sigma_b^2}\pi_{it}(\theta))\right]_{1\le k\le K},$$

$$h_{it} = G_{it}(y_{it} - \pi_{it}(\theta)) - \Phi C_{it}(Y_{it} - X'_{it}\omega)^2,$$

$Vh = \sum_{i=1}^{m}\sum_{t=1}^{n} \mathrm{Var}(h_{it})$, $R = \mathrm{tr}((C\mathbf{V})^2) - \sum_i \sum_t (C\mathbf{V})_{it}^2$ and

$$\Psi_N = (N)^{-1}[Vh + 2\Phi R\Phi']. \tag{22}$$

Note that, for $N \times N$ matrixes C and $C\mathbf{V}$, C_{it} and $(C\mathbf{V})_{it}$ indicate the $((i-1)n+t)$-th diagonal elements respectively.

THEOREM 3. *Suppose Ψ_N in (22) converges to a limiting value Ψ. Under the assumptions A1-A8, as $N \to \infty$, the asymptotic distribution of the goodness-of-fit statistics (21) is*

$$\hat{\chi}^2 \to_d \sum_{j=1}^{K} \lambda_i \mathbb{Z}_i^2, \tag{23}$$

where $\Gamma = diag(\lambda_1, \cdots, \lambda_K)$, and λ_i are the eigenvalues of Ψ and $\mathbb{Z}_1, \cdots, \mathbb{Z}_K$ are i.i.d. $\mathcal{N}(0,1)$.

Let $\hat{\Psi} = N^{-1}[\widehat{Vh} + 2\hat{\Phi}\hat{R}\hat{\Phi}']$ denote the estimate of (22). Computation of $\hat{\Psi}$ is essential to obtaining the critical values in the goodness-of-fit test. In practice, it is often straightforward to evaluate $\hat{\Psi}$ by Monte-Carlo method as follows:

$$\begin{aligned}
\hat{\Psi} &= N^{-1}[\textstyle\sum_i \sum_t \widehat{\mathrm{Var}(h_{it})} + 2\hat{\Phi}\hat{R}\hat{\Phi}'] \\
&\approx N^{-1}\left[\textstyle\sum_i \sum_t \frac{1}{U}\sum_{u=1}^{U}(\hat{h}_{it,(u)} - \overline{h_{it}})(\hat{h}_{it,(u)} - \overline{h_{it}})' + 2\frac{1}{U}\sum_{u=1}^{U}[\hat{\Phi}_{(u)}\hat{R}_{(u)}\hat{\Phi}'_{(u)}]\right] \\
&= \frac{1}{U}\left[N^{-1}\textstyle\sum_i \sum_t (\hat{h}_{it,(u)} - \overline{h_{it}})(\hat{h}_{it,(u)} - \overline{h_{it}})' + 2\frac{1}{N}\sum_{u=1}^{U}\hat{\Phi}_{(u)}\hat{R}_{(u)}\hat{\Phi}'_{(u)}\right],
\end{aligned}$$

where U are the number of Monte-Carlo simulations, $\hat{h}_{it,(l)}$, $\hat{\Phi}_{(l)}$, $\hat{I}^N_{(l)}$, $\hat{R}_{(l)}$ are estimates with θ replaced by $\hat{\theta}$, $\overline{h_{it}} = \frac{1}{U}\sum_{u=1}^{U}\hat{h}_{it,(u)}$, y_{it} are a sample from Bernoulli trials with probability following the fitted BTSM model and β_i are i.i.d. variables generated from $\mathcal{N}(\hat{b}, \hat{\sigma}_b)$. As mentioned in Section 3.2, Laplace's method can be applied to approximate integration in $\frac{\partial \pi_{it}(\theta)}{\partial \omega'}$ and $\frac{\partial \pi_{it}(\theta)}{\partial \sigma_b^2}$.

4.4.2 Finite-Sample Performance and Empirical Application

To examine the finite-sample performance of the proposed tests, we carry out some simulations under nulls and alternatives. Each result is calculated based on 5000

Table 14: BTSM models with four different time series structures

BTSM–	Model	β_i
AR(1)	$logit(\mu_{it}) = \beta_i + 1.3y_{it-1} + 0.3x_{it}$, $x_{it} = 0.2, 0.4, 0.6, 0.8$	$\mathcal{N}(-0.3, 0.5)$
MA(1)	$logit(\mu_{it}) = \beta_i + 1.3(y_{it-1} - \mu_{it-1}) + 0.3x_{it}$, $x_{it} \in (0, 1)$	$\mathcal{N}(-0.3, 0.5)$
AR(2)	$logit(\mu_{it}) = \beta_i + y_{it-1} + 0.5y_{it-2} + 0.3x_{it}$, $x_{it} \in (0, 1)$	$\mathcal{N}(-1, 0.5)$
ARMA(1,1)	$logit(\mu_{it}) = \beta + 1.5y_{it-1} + 0.5(y_{it-1} - \mu_{it-1}) - 0.5x_{it}$, $x_{it} \in (0, 1)$	$\mathcal{N}(-0.3, 0.5)$

simulations with 5% significant level. Two sample sizes N=480 (m=25, n=20) and N=160 (m=16, n=10) and four different partitions (K=2, 4, 6, 8) are studied. For simplicity, we only focus on equal cell partitions in this simulation study. As mentioned in Section 4.1, when unknown parameters are involved, there is no existing test which has valid asymptotic distribution. Thus, we compare our method with a naive test, namely, the Pearson χ^2-test in (20) but with parameters estimated. Since parameters are not assumed to be known, a naive way to apply the Pearson χ^2-test is to modify the asymptotic χ^2 distribution with $K - 1 - a$ degrees of freedom, where a is the number of parameters being estimated. Here, the comparison is conducted only for K=8. For example, for the second model (BTSM-AR(2)) in Table 14, there are five parameters being estimated (three fixed effects, one random effect, and one corresponding variance). As noted above, the χ^2 distribution with $2(= 8 - 1 - 5)$ degrees of freedom may be incorrect (even asymptotically). Since this naive critical value is too small, it is possible that use of the correct critical value would correct the empirical levels in our simulations, but clearly this would only come at the expense of reducing the power further.

Binary data are generated by using the BTSM models listed in Table 14 with four different time series structures. Table 15 reports the empirical rejection probabilities associated with these four models to examine the empirical level of the test. In general, when the sample size increases, the empirical level of the proposed test becomes more stable with respect to the number of partitions K. Compared with the naive

Table 15: Empirical level of the goodness-of-fit test at 5 %

Model	(m, n)	$K = 2$	$K = 4$	$K = 6$	$K = 8$	χ^2
BTSM–AR(1)	$(24, 20)$	0.042	0.046	0.046	0.044	0.094
	$(16, 10)$	0.047	0.049	0.054	0.055	0.119
BTSM–MA(1)	$(24, 20)$	0.046	0.052	0.054	0.062	0.084
	$(16, 10)$	0.063	0.056	0.065	0.074	0.172
BTSM–AR(2)	$(24, 20)$	0.056	0.042	0.048	0.06	0.151
	$(16, 10)$	0.06	0.048	0.054	0.062	0.277
BTSM–ARMA(1,1)	$(24, 20)$	0.044	0.046	0.042	0.046	0.294
	$(16, 10)$	0.055	0.058	0.051	0.056	0.346

Pearson χ^2-test, the proposed method performs better in the following two respects. First, when the number of estimated parameters involved in the model increases, the proposed method provides a more stable empirical level. For example, the empirical level of the naive test almost doubles and far exceeds the nominal 5% level when the number of estimated parameters increases from four (AR(1) or MA(1)) to five (AR(2) or ARMA(1,1)). This is because the critical value of the naive test decreases rapidly when the number of estimated parameters increases. The other advantage of the proposed method is the performance robustness to sample size. For the naive test, the empirical level increases dramatically when the sample size decreases, while for the proposed method, the increase is slight to modest.Table 16 reports the computing times on a 3.4-GHz PC for calculating the empirical level (based on 5000 simulations) using R. The computing time increases linearly with sample size, while it increases marginally with K.

In terms of power, we choose two types of alternatives to assess the distributional assumptions involved in the fitted model (at 5% level), including the Bernoulli assumption for the binary data and the normal assumption for the random effects. The first alternative assumes that the random effects are normally distributed and the binary data follow a *beta-binomial* distribution. That is, y_{it} are generated from Bernoulli($\mathcal{P}_{it}$) distribution, and $\mathcal{P}_{it}$ is a random variable with a Beta(μ_{it}, $1 - \mu_{it}$)

Table 16: Computing times (in minutes) for calculating empirical level

Model	(m, n)	$K = 2$	$K = 4$	$K = 6$	$K = 8$	χ^2
BTSM–AR(1)	$(24, 20)$	13	15	15	17	17
	$(16, 10)$	5	6	6	6	6
BTSM–MA(1)	$(24, 20)$	16	18	18	20	20
	$(16, 10)$	7	7	7	9	9
BTSM–AR(2)	$(24, 20)$	18	18	20	20	20
	$(16, 10)$	6	7	7	8	8
BTSM–ARMA(1,1)	$(24, 20)$	18	18	21	22	22
	$(16, 10)$	7	7	8	8	8

distribution. The other alternative assumes a departure from the normal assumption for random effects. Let y_{it} follow Bernoulli(μ_{it}) distribution, and the random effect $\boldsymbol{\beta}$ follows a mixture of two normal distributions $\mathcal{N}(\mathbf{b}_1, 1)$ and $\mathcal{N}(\mathbf{b}_2, 1)$ with probability *prob* and $1 - prob$, denoted by $MIXN(\mathbf{b}_1, \mathbf{b}_2, prob)$. In the simulation, the random effect is assume to be $MIXN(-\mathbf{0.5}, \mathbf{0.5}, 0.3)$. For each alternative, μ_{it} are obtained from the values specified in the four models given in Table 14. Based on the generated data, models are fitted by the procedure described in Section 3.

Tables 17 and 18 report the empirical rejection probability for both alternatives associated with four BTSM models to examine the empirical power. Their corresponding computing times are listed in Tables 19 and 20. Clearly, for the first two models, the proposed test is more powerful than the naive test for both alternatives (with the exception of BTSM–AR(1), $m = 24$, $n = 20$, $K = 2$ in Table 18). In some cases of the last two models, when K is small (mostly for $K = 2$, and some for $K = 4$), the naive method has more power than the proposed method, but this is due to the higher empirical levels of the former in Table 15. Another issue is the dependence of performance on K, the number of cells. It is well known that the power of this type of goodness-of-fit test can vary greatly with K. This is observed in the simulation results, especially when the sample size is smaller. Therefore, proper choice of partitions is important. This leads to the following guidelines for choosing

Table 17: Power of testing Bernoulli assumption under beta-binomial distribution

Model	(m, n)	$K = 2$	$K = 4$	$K = 6$	$K = 8$	χ^2
BTSM–AR(1)	$(24, 20)$	0.651	0.659	0.718	0.772	0.574
	$(16, 10)$	0.361	0.403	0.608	0.548	0.296
BTSM–MA(1)	$(24, 20)$	0.792	0.806	0.811	0.912	0.778
	$(16, 10)$	0.528	0.729	0.579	0.634	0.428
BTSM–AR(2)	$(24, 20)$	0.688	0.858	0.762	0.756	0.728
	$(16, 10)$	0.603	0.596	0.586	0.594	0.578
BTSM–ARMA(1,1)	$(24, 20)$	0.402	0.818	0.754	0.746	0.648
	$(16, 10)$	0.216	0.286	0.428	0.502	0.486

Table 18: Power of testing normal random effect under mixed normal distribution

Model	(m, n)	$K = 2$	$K = 4$	$K = 6$	$K = 8$	χ^2
BTST–AR(1)	$(24, 20)$	0.319	0.427	0.486	0.674	0.354
	$(16, 10)$	0.205	0.314	0.458	0.327	0.135
BTSM–MA(1)	$(24, 20)$	0.994	1	1	0.998	0.993
	$(16, 10)$	0.924	0.983	0.988	0.97	0.826
BTSM–AR(2)	$(24, 20)$	0.596	0.607	0.686	0.702	0.618
	$(16, 10)$	0.336	0.375	0.395	0.448	0.432
BTSM–ARMA(1,1)	$(24, 20)$	0.546	0.658	0.586	0.616	0.518
	$(16, 10)$	0.323	0.348	0.356	0.434	0.346

the optimal number of partitions.

Although the construction of the goodness-of-fit test allows arbitrary partition of the cells, its performance depends on a proper choice of the number of exclusive subsets K in (19). How to choose the optimal number of partitions? First, to ensure enough power, K should not be too small, because the fewer cells the more difficult to distinguish between two distributions. On the other hand, if there are too many cells, the size of the test may become a problem. This is because the asymptotic distribution of the test is based on a K-dimensional central limit theorem. A necessary condition to maintain this asymptotic property is that $K/N^{1/5} \to 0$ (Senatov, 1980; Jiang, 2001b). Therefore, the proper number of partitions should be chosen from 1 to $\lceil N^{1/5} \rceil$. Within this range, conducting a simulation with comparable sample size will

Table 19: Computing times (in minutes) for Table 17

Model	(m, n)	$K = 2$	$K = 4$	$K = 6$	$K = 8$	χ^2
BTSM–AR(1)	$(24, 20)$	9	9	13	14	14
	$(16, 10)$	2	2	2	3	3
BTSM–MA(1)	$(24, 20)$	9	9	12	12	12
	$(16, 10)$	3	3	4	4	4
BTSM–AR(2)	$(24, 20)$	11	11	11	13	13
	$(16, 10)$	5	6	6	6	6
BTSM–ARMA(1,1)	$(24, 20)$	10	11	15	16	16
	$(16, 10)$	5	5	5	6	6

Table 20: Computing times (in minutes) for Table 18

Model	(m, n)	$K = 2$	$K = 4$	$K = 6$	$K = 8$	χ^2
BTSM–AR(1)	$(24, 20)$	9	9	10	10	10
	$(16, 10)$	5	5	5	5	5
BTSM–MA(1)	$(24, 20)$	11	11	11	13	13
	$(16, 10)$	5	6	6	6	6
BTSM–AR(2)	$(24, 20)$	8	9	9	11	11
	$(16, 10)$	4	5	5	5	5
BTSM–ARMA(1,1)	$(24, 20)$	11	11	13	13	13
	$(16, 10)$	5	5	6	7	7

be helpful in determining the optimal number of partitions. R code for the simulations are available on http://www2.isye.gatech.edu/∼jeffwu/publications/, which can be easily implemented.

4.5 Application in Adhesion Frequency Experiment

In this section, we revisit the adhesion frequency experiment data and apply the proposed model to predict the adhesion probability. As in Section 2, there are 37 pairs of cells used in this experiment. Adhesion test cycles for each pair are repeated 50 times. To study the time series behavior, for every subject, the first five observations are treated as additional predictor variables. Therefore, in this example, m=37, n=45. The covariate here is the average number of bonds denoted by ANB_i for the i-th pair of cells. For each pair of cells, the ANB is fixed. Therefore, there is no time-dependent covariate in this example and the one-dimensional (p=1) covariates in model (7) can be simplified by assuming $x_{it} = x_{it,1} = ANB_i$, for all t.

With fixed effects

$$\omega = (\alpha_1, \gamma_1, \varphi_1, \zeta_1),$$

and the corresponding $X'_{it} = (ANB_i, ANB_i \times y_{it-1}, y_{i\,t-1}, (y_{i\,t-1} - \mu_{i\,t-1}))$, the fitted BTSM model for adhesion probability is given below:

$$logit(\mu_{it}) = \beta_i + \alpha_1 ANB_i + \gamma_1 ANB_i \times y_{it-1} + \varphi_1 y_{it-1} + \zeta_1 (y_{i\,t-1} - \mu_{i\,t-1}), \qquad (24)$$

where $\beta_i \sim \mathcal{N}(-1.33, 0.44)$. The value of the MQPLE is

$$\hat{\omega} = (0.97, -0.62, 1.76, -0.86)$$

with the corresponding p-values 0.004, 0.031, $<$ 0.001, and 0.006. The estimated variance component $\hat{\sigma}_b$=0.4 (with standard deviation 0.14) provides clear evidence on the substantial heterogeneity among subjects. In model (24), ANB_i and $y_{i\,t-1}$ have significant effects on the cell adhesion probability at time t. The positive α_1 value of 0.97 indicates that the cell adhesion probability increases with respect to the

ANB. The adhesion memory can be described by a first-order autoregressive and moving average process. The positive φ_1 value of 1.76 indicates that the adhesion probability is higher if adhesion occurs in the previous test. The significant interaction ($ANB_i \times y_{it-1}$) plays an important role in the model interpretation. Based on the fitted model (24), the coefficient of the ANB_i, $0.97 - 0.62y_{it-1}$, shows that the effect of ANB is smaller if an adhesion occurs in the previous test (y_{it-1}=1). On the other hand, based on the coefficient of y_{it-1}, i.e., $1.76 - 0.86 - 0.62ANB_i$=$0.9 - 0.62ANB_i$, the effect of y_{it-1} is reduced as the average number of bonds increases. Furthermore, the memory effect is close to 0 if the ANB is around 1.45 (=0.9/0.62). It implies that, if two cells with ANB more than 4.3 seconds, the repeated adhesion tests become nearly independent. This model gives so much new information on the adhesion frequency analysis, because it provides not only a flexible model for considering the memory effect but also the conditions under which the independence assumption may hold.

The distributional assumptions here are the normally distributed random effects and the dependent Bernoulli distributed responses. To assess their adequacy, the proposed goodness-of-fit test (21) is applied in this example. Based on some simulation studies that we suggested in section 4.2, the optimal number of partition in this example is $K = 4$. Therefore, we first partition the the previous information space H_{it} into four disjoint events as follows:

$$E_1 = (y_{it-1} = 0, \mu_{it-1} > 0.5), \quad E_2 = (y_{it-1} = 0, \mu_{it-1} \leq 0.5),$$
$$E_3 = (y_{it-1} = 1, \mu_{it-1} > 0.5), \quad E_4 = (y_{it-1} = 1, \mu_{it-1} \leq 0.5).$$

That is, c_1=1, c_2=2, c_3=2 in (19). 5000 Monte-Carlo simulations are conducted to evaluate $\hat{\Psi}$. The corresponding eigenvalues for $\hat{\Psi}$ are {0.3100, 0.1428, 0.0350, 0.0252}. By Theorem 3, the critical values of the proposed goodness-of-fit test at α=0.01, 0.05, and 0.1 are 2.3819, 1.6271, and 1.2556 respectively. The test statistic under model (24) is $\hat{\chi}^2$=0.9392, which is much smaller than the critical values. There is thus no

evidence to reject the hypothesis that the binary responses in adhesion tests follow a dependent Bernoulli distribution with probability given by model (24). Similar to the study in Section 4.2, we compare the proposed test with the naive χ^2 test in (20) with one degree of freedom. The nave test statistic has the value 6.8838 with the corresponding p-value of 0.0087. This would lead to the rejection of the hypothesis of dependency and the model in (21). In view of the simulation results in Section 4.2 that the naive test can have an exceedingly large test statistic value, such a conclusion cannot be taken seriously.

Recall that the preliminary analysis in Section 2 shows some memory effects in the repeated observations. By applying the BTSM model, the cell adhesion memory can be described by an ARMA(1,1) process. Besides, model (24) can quantify the effect of average number of bonds and identify a significant interaction between average number of bonds and the previous test result. This is a great advantage, because in practice it is difficult to assess the moving average and interaction effects by graphical analysis. As shown in this example, by including the interaction term, the BTSM model provides flexibility in capturing different time series structures with respect to different covariates. Given the fitted models, goodness-of-fit tests are conducted to check the distributional assumptions. The test result provides statistical evidence on the adequacy of the distributional assumption and supports model-based predictions. Another advantage of the BTSM model is that it incorporates the random effects. Thus inference and predictions can be made beyond the particular subjects used in the experiment.

4.6 Summary and Concluding Remarks

Despite the prevalence of multiple binary time series data in many applications, their modeling and inference have not been systematically studied in the literature. We propose a binary time series mixed model (BTSM) to analyze data when a repeated

binary time series is observed for each subject. It handles multiple time series by incorporating random effects to borrow strength across different subjects. Thus, inference and predictions can be made beyond the specific units in the study. The BTSM model includes numerous known models as special cases. Moreover, it may have applications in longitudinal analysis.

Estimators for the fixed effects and variance components are shown to be consistent and asymptotically normally distributed. To assess the adequacy of the distributional assumptions in the BTSM model, we propose a new goodness-of-fit test. Because there are some unknown parameters and the data are dependent, the asymptotic distribution for the test statistic is derived by using a martingale central limit theorem. Not surprisingly, the results are different from the classical Pearson's χ^2 test. The proposed test outperforms the naive Pearson's χ^2 test in an simulation study. Some guidelines are given on the choice of the optimal number K of partitions.

As an application, the BTSM model is applied to fit some multiple binary time series observed on T-cell adhesion frequency experiment. This study demonstrates how the BTSM model can help in quantitatively describing the effects of significant factors. Furthermore, the fitted model provides valuable information on moving average and interaction effects, which cannot be obtained from graphical analysis. This example shows that the first-order autocorrelation effect can be observed from graphical analysis, but not when higher order autocorrelations are present. The goodness-of-fit test is also demonstrated in this example. Although the covariates in this example are independent of time, the proposed model and inference are generally applicable to problems with time-dependent covariates.

APPENDIX A

PROOF OF LEMMA 1

Since each column in the $LHD(n,k)$ is a permutation of $\{1,2,\cdots,n\}$, we have

$$\begin{aligned}\textstyle\sum_{i=1}^{\binom{n}{2}} d_i = & \textstyle\sum_{i=2}^{n}\sum_{j=1}^{i-1} d(\mathbf{s_i},\mathbf{s_j}) \\ = & \textstyle\sum_{l=1}^{k}\sum_{i=2}^{n}\sum_{j=1}^{i-1}|s_{il}-s_{jl}| \\ = & \textstyle k\sum_{i=2}^{n}\sum_{j=1}^{i-1}|s_{i1}-s_{j1}|.\end{aligned}$$

Without loss of generality, we can take the first column as $(1,2,\cdots,n)'$. Therefore,

$$\begin{aligned}\textstyle\sum_{i=1}^{\binom{n}{2}} d_i = & \textstyle k\sum_{i=2}^{n}\sum_{j=1}^{i-1}|i-j| \\ = & \textstyle k\sum_{i=2}^{n}\frac{i(i-1)}{2} \\ = & \frac{kn(n^2-1)}{6}.\end{aligned}$$

Thus,

$$\begin{aligned}\overline{d} = & \frac{kn(n^2-1)/6}{\binom{n}{2}} \\ = & \frac{(n+1)k}{3}\ .\end{aligned}$$

APPENDIX B

PROOF OF LEMMA 2

For $m = 2$,

$$\frac{1}{\sum_{j=1}^{k} d_{j1}} + \frac{1}{\sum_{j=1}^{k}(c_j - d_{j1})} = \frac{\sum_{j=1}^{k} c_j}{\sum_{j=1}^{k} d_{j1} \times \sum_{j=1}^{k}(c_j - d_{j1})},$$

where $c_j = d_{j1} + d_{j2}$, for all $j = 1, 2, \cdots, k$. Since $\sum_{j=1}^{k} c_j$ is a constant, it is easy to see that the right side is a maximum when $\sum_{j=1}^{k} d_{j1} = \sum_{j=1}^{k} d_{j(1)}$. Therefore,

$$\sum_{i=1}^{2} \frac{1}{\sum_{j=1}^{k} d_{ji}} \leq \sum_{i=1}^{2} \frac{1}{\sum_{j=1}^{k} d_{j(i)}}.$$

Thus, the result holds for $m = 2$. Assume the upper bound is achieved by the ordered sequence for $m = M$. When $m = M + 1$, suppose the upper bound is achieved by some unordered sequence $\{d_{j1^*}, \cdots, d_{jM+1^*}\}$. So the upper bound is $\sum_{i=1}^{M+1} \frac{1}{\sum_{j=1}^{k} d_{ji^*}}$. Without loss of generality, assume that at least the first sequence does not follow the order. Because of this, there always exists an M-element subset $\{d_{11^*}, \cdots, \widehat{d_{1t^*}} \cdots, d_{1M+1^*}\}$ that does not follow the order, where the notation $\widehat{d_{1t^*}}$ means that the sequence is without d_{1t^*}. But since the upper bound holds for $m = M$, we have

$$\frac{1}{\sum_{j=1}^{k} d_{j1^*}} + \cdots + \widehat{\frac{1}{\sum_{j=1}^{k} d_{jt^*}}} \cdots + \frac{1}{\sum_{j=1}^{k} d_{jM+1^*}} \leq \frac{1}{\sum_{j=1}^{k} d_{j(1)}} + \cdots + \widehat{\frac{1}{\sum_{j=1}^{k} d_{jt^*}}} \cdots + \frac{1}{\sum_{j=1}^{k} d_{j(M+1)}}.$$

This is a contradiction, because by adding $1/\sum_{j=1}^{k} d_{jt^*}$ to both sides we obtain

$$\sum_{i=1}^{M+1} \frac{1}{\sum_{j=1}^{k} d_{ji^*}} \leq \frac{1}{\sum_{j=1}^{k} d_{j(1)}} + \cdots + \frac{1}{\sum_{j=1}^{k} d_{jt^*}} \cdots + \frac{1}{\sum_{j=1}^{k} d_{j(M+1)}},$$

which is a better upper bound. By mathematical induction, we can prove that the function achieves the upper bound when all k sequences are in increasing order.

APPENDIX C

PROOF OF PROPOSITION 1

To find a lower bound for ϕ_p, consider the following minimization problem with respect to $d_1, \cdots, d_{\binom{n}{2}}$.

$$\min \phi_p = \left(\sum_{i=1}^{\binom{n}{2}} \frac{1}{d_i^p} \right)^{1/p} \quad \text{subject to} \quad \sum_{i=1}^{\binom{n}{2}} d_i = \binom{n}{2}\bar{d},$$

where $\bar{d} = (n+1)k/3$. Using the Lagrange multiplier method, it is easy to show that the optimal solution is given by $d_1 = d_2 = \cdots = d_{\binom{n}{2}} = \bar{d}$. Therefore, $\binom{n}{2}^{1/p}/\bar{d}$ is a lower bound for ϕ_p. But since we know the d_i's in an LHD are integers, a better lower bound can be obtained by adding this constraint to the above optimization problem. To find the optimal solution under the integer restriction, consider the following two groups: $I = \{i : d_i \leq \bar{d}\}$ and $II = \{i : d_i > \bar{d}\}$. Since the sum of the d_i's is a constant, if we increase a d_i for an $i \in I$, then we should decrease a d_i, $i \in II$, by the same amount . It is easy to show that such a change will decrease ϕ_p. Therefore, the minimum of ϕ_p can be achieved by

$$d_1 = \cdots = d_N = \lfloor \bar{d} \rfloor \ ; \ d_{N+1} = \cdots = d_{\binom{n}{2}} = \lceil \bar{d} \rceil,$$

provided such an N exists. We must have $N\lfloor \bar{d} \rfloor + \{\binom{n}{2} - N\}\lceil \bar{d} \rceil = \binom{n}{2}\bar{d}$, which gives $N = \binom{n}{2}(\lceil \bar{d} \rceil - \bar{d})$. This is a feasible solution, because $\binom{n}{2}\bar{d} = (n+1)k$ is an integer. Thus

$$\phi_p \geq \left(\frac{N}{\lfloor \bar{d} \rfloor^p} + \frac{\binom{n}{2} - N}{\lceil \bar{d} \rceil^p} \right)^{1/p} = \phi_{p,L}.$$

Now consider the upper bound. All the k factors have the same inter-site distances $\{d_{j,1}, \cdots, d_{j,\binom{n}{2}}\}$, where $j = 1, \cdots, k$. For example, if $n = 5$, the inter-site distances

for each factor is $\{1,1,1,1,2,2,2,3,3,4\}$. In general, $(n-1)$ of the $d_{j,i}$'s are 1, $(n-2)$ of the $d_{j,i}$'s are $2,\cdots$, and one is $(n-1)$. Different LHDs have different combinations of the inter-site distances of each factor. Therefore $d_i = \sum_{j=1}^{k} d_{j,i}$, where $i = 1,\cdots,\binom{n}{2}$. By Lemma 2

$$\phi_p = \left(\sum_{i=1}^{\binom{n}{2}} \frac{1}{d_i^p}\right)^{1/p} = \left(\sum_{i=1}^{\binom{n}{2}} \frac{1}{\sum_{j=1}^{k} d_{j,i}^p}\right)^{1/p} \leq \left(\sum_{i=1}^{\binom{n}{2}} \frac{1}{\sum_{j=1}^{k} d_{j,(i)}^p}\right)^{1/p}$$

Note that the inter-site distances of each of the k factors is ordered in the same way. Therefore, $(n-1)$ of the d_i's are k, $(n-2)$ of the d_i's are $2k,\cdots$, and one is $(n-1)k$. Thus

$$\phi_p \leq \left(\sum_{i=1}^{\binom{n}{2}} \frac{1}{\sum_{j=1}^{k} d_{j,i}^p}\right)^{1/p} = \left\{\sum_{i=1}^{n-1} \frac{(n-i)}{(ik)^p}\right\}^{1/p} = \phi_{p,U}.$$

APPENDIX D

BAYESIAN VARIABLE SELECTION TECHNIQUE

Here we provide some additional details for the Bayesian variable selection technique. The computer model can be represented as $Y = f(\mathbf{x})$, where the transfer function f can be highly nonlinear. First assume that each x_i takes only three values 1, 2, and 3. Later we will explain how to generalize this. Define the linear and quadratic effects for each x_i as in Eq. (6). Now consider approximating $f(\mathbf{x})$ by a linear model containing all of the interaction terms (up to the pth order interaction). The linear model can be written as $\sum_{i=0}^{3^p-1} \beta_i u_i$, where $u_0 = 1, u_1 = x_{1l}, \ldots$, and $u_{3^p-1} = x_{1q} \cdots x_{pq}$.

A major step in the Bayesian variable selection technique is to postulate a prior distribution for $\beta = (\beta_{\mathbf{0}}, \ldots, \beta_{\mathbf{3^p-1}})'$. This is a difficult task because of the huge number of parameters. To simplify this task, Joseph (2006b) and Joseph and Delaney (2007) proposed an interesting idea. Instead of directly postulating a prior for β, postulate a functional prior for $f(\mathbf{x})$ and use it to induce a prior for β. Assume that

$$f(\mathbf{x}) \sim \mathbf{GP}(\mu_{\mathbf{0}}, \sigma_{\mathbf{0}}^{\mathbf{2}}\psi),$$

where μ_0 is the mean and $\sigma_0^2\psi$ is the covariance function of the Gaussian process (GP). Because there are 3^p parameters in the linear model, their distribution can be obtained based on 3^p function values. One simple choice is to evaluate the function at the full factorial design for the p factors (which contains 3^p points). To simplify the results further, write the linear model as $\mu_0 + \sum_{i=0}^{3^p-1} \beta_i u_i$ and assume a product correlation structure given by $\psi(\mathbf{h}) = \prod_{\mathbf{j=1}}^{\mathbf{p}} \psi_{\mathbf{j}}(\mathbf{h_j})$. Then, it can be shown that

(Joseph and Delaney (2007))

$$\begin{aligned}
\beta_0 &\sim \mathcal{N}\left(0, \tau_0^2\right), \\
\beta_1 &\sim \mathcal{N}\left(0, \tau_0^2 r_{1l}\right), \\
\beta_2 &\sim \mathcal{N}\left(0, \tau_0^2 r_{1q}\right), \\
&\vdots \\
\beta_{3^p-1} &\sim \mathcal{N}\left(0, \tau_0^2 r_{1q} r_{2q} \cdots r_{pq}\right),
\end{aligned}$$

where r_{jl} and r_{jq} for $j = 1, \cdots, p$ are calculated using Eq. (7). Further, Joseph and Delaney (2007) shows that β_i's are approximately independent. Thus, the prior distribution for β is a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\tau_0^2\mathbf{R} = \mathbf{diag}\{\mathbf{1}, \mathbf{r_{1l}}, \mathbf{r_{1q}}, \ldots, \mathbf{r_{1q}} \cdots \mathbf{r_{pq}}\}$.

Let the experiment has n runs and let $\mathbf{y}$ be the data. We have $\mathbf{y} = \mu_\mathbf{0}\mathbf{1} + \mathbf{U}\beta$, where $\mathbf{U}$ is the model matrix with dimension $n \times 3^p$. Using Bayes theorem, the posterior distribution of β is given by

$$\beta|\mathbf{y} \sim \mathcal{N}\left(\frac{\tau_\mathbf{0}^\mathbf{2}}{\sigma_\mathbf{0}^\mathbf{2}}\mathbf{R}\mathbf{U}'\mathbf{\Psi}^{-\mathbf{1}}(\mathbf{y} - \mu_\mathbf{0}\mathbf{1}),\ \tau_\mathbf{0}^\mathbf{2}\mathbf{R} - \frac{\tau_\mathbf{0}^\mathbf{4}}{\sigma_\mathbf{0}^\mathbf{2}}\mathbf{R}\mathbf{U}\mathbf{\Psi}^{-\mathbf{1}}\mathbf{U}\mathbf{R}\right).$$

The posterior mean can be used as an estimate of β.

Note that if we are interested only up to the two-factor interactions, then approximate results can be obtained by replacing $\mathbf{R}$ and $\mathbf{U}$ by their appropriate sub-matrices. Moreover, if a factor takes values in a continuous interval, then it should be scaled in the interval $[1.0, 3.0]$. Other ranges such as $[0, 1]$ or $[-1, 1]$ may also be used. However, the formulas for the linear-quadratic effects and r's should be changed accordingly. For example, if the factors are scaled in $[0, 1]$, then the linear-quadratic effects should be calculated using Eq. (6) after replacing $x_j - 2$ with $2(x_j - .5)$ and the r's using Eq. (7) after replacing the arguments of ψ_j by .5 and 1 instead of 1 and 2.

APPENDIX E

PROOF OF PROPOSITION 2

If there is one branching factor ($q = 1$), a BLHD will include k_1 small LHD(n_1, m_1) for the k_1 levels of the branching factor and a LHD(N, t) for the shared factors. ϕ_λ can be written as

$$\phi_\lambda = \left(\sum_{\boldsymbol{g} \neq \boldsymbol{h}} \left[\frac{t}{d_x(\boldsymbol{g}, \boldsymbol{h})} \right]^\lambda + \sum_{i=1}^{k_1} \sum_{g_{\delta_1} = h_{\delta_1} z_{1,i}} \left[\frac{m_1 + t}{d_{v_1}(\boldsymbol{g}, \boldsymbol{h}) + d_x(\boldsymbol{g}, \boldsymbol{h})} \right]^\lambda \right)^{1/\lambda}$$

As shown in Joseph and Hung (2008), for a given LHD(N,t), the average inter-site distance (rectangular measure) is $N(N^2-1)t/6$, which is a constant. With these constraints, finding a lower bound for ϕ_λ can be formulated as a constraint minimization problem:

$$\begin{aligned} &\min \phi_\lambda \\ \text{subject to} \quad &\textstyle\sum_{g_{\delta_1} = h_{\delta_1} = z_{1,i}} d_{v_1}(\boldsymbol{g}, \boldsymbol{h}) = \frac{m_1 n_1 (n_1^2 - 1)}{6}, 1 \leq i \leq k_1 \\ &\textstyle\sum_{\boldsymbol{g} \neq \boldsymbol{h}} d_x(\boldsymbol{g}, \boldsymbol{h}) = \frac{tN(N^2-1)}{6}, \end{aligned} \tag{25}$$

where $\sum_{g_{\delta_1} = h_{\delta_1} = z_{1,i}} d_{v_1}(\boldsymbol{g}, \boldsymbol{h})$ is the sum of inter-site distances for those smaller LHD(n_1,m_1) and $\sum_{\boldsymbol{g} \neq \boldsymbol{h}} d_x(\boldsymbol{g}, \boldsymbol{h})$ is the sum of inter-site distances for LHD(N,t). Since

$$\phi_\lambda \geq \phi_\lambda^* = \left(\sum_{g_{\delta_1} \neq h_{\delta_1}} \left[\frac{t}{d_x(\boldsymbol{g}, \boldsymbol{h})} \right]^\lambda + \sum_{i=1}^{k_1} \sum_{g_{\delta_1} = h_{\delta_1} = z_{1,i}} 2 \left[\frac{m_1 + t}{d_{v_1}(\boldsymbol{g}, \boldsymbol{h}) + d_x(\boldsymbol{g}, \boldsymbol{h})} \right]^\lambda \right)^{1/\lambda},$$

the lower bound can be found by minimizing ϕ_λ^* with the same constraints in (25). Hence, the lower bound can be solved by using the Lagrange multiplier method. For the upper bound of ϕ_λ, result for BLHD is a simple extension of that for LHD. Therefore, they can be proved by the same argument used in Joseph and Hung (2008).

APPENDIX F

ASSUMPTIONS

A1. The parameter ω *belongs to an open set* $B \subseteq R^s$.

A2. The covariate matrix X_{it} *lies almost surely in a nonrandom compact subset of* R^s *such that* $P[\sum_i \sum_t X'_{it} X_{it} > 0] = 1$.

A3. $\sigma_b \geq 0$ *and* $Var(\beta_1^2) > 0$.

A4. As $N \to \infty$, $lim\, inf\, \lambda_{min} Cor(I_{N-s}, \mathbf{V}_1) > 0$ *and* $lim\, tr(\mathbf{V}'_1 \mathbf{V}_1)^{1/2} = \infty$, *where for matrices* $A_1, A_2, Cor(A_1, A_2) = tr(A'_1 A_2)/[tr^{1/2}(A'_1 A_1) tr^{1/2}(A'_2 A_2)]$.

A5. $\|N^{-1/2} a'_T D\|$ *and* $\|(N)^{-1/2} a'_T \Phi J^{1/2}\|$ *are bounded, where* $\|\kappa\| = (\kappa'\kappa)^{1/2}$ *for any vector* κ.

A6. Define $\Delta' = (\mathbf{Y} - \mathbf{X}\omega)' \mathbf{V}^{-1/2}$ *as a vector with elements* Δ_{it}, *and* $\sigma_N^2 = \sum_i \sum_t Var\Big((C^* \mathbf{V})_{it} \Delta_{it}^2 + G^*_{it}(y_{it} - \pi_{it}(\theta)) \Big) + 2 \sum_{(it)' \neq it} (C^* \mathbf{V})^2_{it,(it)'}$. *There exists numbers* L_{it}, *such that as* $N \to \infty$, *the following quantities converge to 0:*

$\sigma_N^{-2} \{ \sum_{it} E[(C^* \mathbf{V})_{it} (\Delta_{it}^2 - 1)]^2 \mathbf{1}_{(|\Delta_{it}| > L_{it})} + \frac{1}{2} \sum_{it \neq (it)'} ((C^* \mathbf{V})_{it,(it)'})^2 [\delta_{it} + \delta_{(it)'}] \}$,

and $\sigma_N^{-2} \sum_{it} (G^*_{it})^2 E(y_{it} - \pi_{it}(\theta))^2 \mathbf{1}_{(|y_{it} - \pi_{it}(\theta)| > L_{it})}$, *where* $\delta_{it} = E\Delta_{it}^2 \mathbf{1}_{(|\Delta_{it}| > L_{it})}$.

A7. There exists numbers L_{it}, *such that as* $N \to \infty$, *the following quantities converge to 0:*

$\sigma_N^{-4} \{ \sum_{it} E[(C^* \mathbf{V})_{it} (\Delta_{it}^2 - 1)]^4 \mathbf{1}_{(|\Delta_{it}| \leq L_{it})} + \sum_{it \neq (it)'} ((C^* \mathbf{V})_{it,(it)'})^4 \delta^*_{it} \delta^*_{(it)'} +$

$\sum_{it} [\sum_{(it)' \neq it} ((C^* \mathbf{V})_{it,(it)'})^2]^2 \delta^*_{it} \}$, *and* $\sigma_N^{-4} \sum_{it} (G^*_{it})^4 E(y_{it} - \pi_{it}(\theta))^4 \mathbf{1}_{(|y_{it} - \pi_{it}(\theta)| \leq L_{it})}$, *where* $\delta^*_{it} = E\Delta_{it}^4 \mathbf{1}_{(|\Delta_{it}| \leq L_{it})}$.

A8. As $N \to \infty$, $\lambda_{max} \xi' \xi / \sigma_N^2 \to 0$, *where* $\xi = (C^* \mathbf{V}) - diag(C^* \mathbf{V})$.

Assumptions A1 and A2 are required for the asymptotic properties for fixed effects estimated from partial likelihood. Lindeberg's condition holds under assumption A2

(Fokianos and Kedem, 1998), which leads to the proof of Theorem 1. Assumptions A3 and A4 are the key conditions for consistency and asymptotic normality of the REML variance components estimator. These two assumptions are the same as in Jiang (1996).

APPENDIX G

PROOF OF THEOREM 1

Based on the partial likelihood, differentiation of (9) with respect to ω leads to the partial score process

$$S_n(\omega, \sigma_b) = \sum_{t=1}^{n} \sum_{i=1}^{m} X_{it}(y_{it} - \pi_{it}(\omega, \sigma_b)).$$

Assume a σ-field is generated from the past data and covariates

$$\mathcal{F}_{n-1} = \sigma(H_{1n}, H_{2n}, \cdots, H_{mn}).$$

It is clear that $\mathrm{E}[S_n(\omega, \sigma_b) \mid \mathcal{F}_{n-1}] = \mathrm{E}[S_{n-1}(\omega, \sigma_b)]$, and $\mathrm{E}[S_n(\omega, \sigma_b)] = 0$. Base on this fact and A1 and A2, it is easy to see that the partial score process $S_n(\omega, \sigma_b)$ is the sum of zero-mean martingale differences with respect to $\mathcal{F}_{n-1}$. The asymptotic normality follows from the martingale central limit theorem (Theorem 3.2 in Hall and Heyde (1980)). Detail of the proof is analogue to Slud and Kedem (1994).

APPENDIX H

PROOF OF THEOREM 2

The inference for the variance component can be formulated as a linear mixed model with variances of error terms following the GLM iterative weights in (17). Define $\mathbf{Y}^* = \mathbf{W}^{1/2}\mathbf{Y}$, $\mathbf{X}^* = \mathbf{W}^{1/2}\mathbf{X}$, $\mathbf{Z}^* = \mathbf{W}^{1/2}\mathbf{Z}$, $\epsilon^* = \mathbf{W}^{1/2}\epsilon$. Replacing them in (17), we have $\mathbf{Y}^* = \mathbf{X}^*\omega + \mathbf{Z}^*\boldsymbol{\beta} + \epsilon^*$, with $\epsilon^* \sim \mathcal{N}(0, I)$ and $\boldsymbol{\beta}$ following the same distribution in (17). The results directly follow as a special case of Theorem 4.1 of Jiang (1996) with the variance component parameter space $\Theta = \{\sigma_b^2 \geq 0\}$.

APPENDIX I

APPENDIX D: PROOF OF THEOREM 3

The proof is along the lines of Jiang (2001b). It consists of several lemmas that culminate in the final proof.

Lemma D.1. *Under the same assumptions in Theorem 3, define*

$$\Psi_n^{(1)} = \Psi_n^{(1)}(\theta) = n^{-1}\sum_{i=1}^{n} Var(h_{n,i}^{(1)}),$$

where

$$h_i^{(1)} = \left[\mathbf{1}_{(H_i\in E_k)} - \left(\frac{1}{n}\sum_{j=1}^{n}(\mathbf{1}_{(H_j\in E_k)}X_j'(1-\pi_j(\omega))\pi_j(\omega))\right)\Lambda_n^{-1}(\omega)X_i\right]_{1\le k\le K}(y_i-\pi_i(\omega)).$$

Suppose $\Psi_n^{(1)}$ converges to a limiting value $\Psi^{(1)}$. If there is no random effect in model (7)(i.e. $m = 1$ in Theorem 3), the asymptotic distribution of the test statistic (21) is

$$\hat{\chi}^2 = \frac{1}{n}\sum_{j=1}^{K}(M_j - e_j(\hat{\omega}))^2 \to_d \sum_{k=1}^{K}\lambda_k\mathbb{Z}_k^2, \tag{26}$$

where $\lambda_1, \cdots, \lambda_K$ are the eigenvalues of $\Psi^{(1)}$.

Proof:

Let $\xi_n = (\xi_{n,k})_{1\le k\le K}$,

$$\xi_{n,k} = M_k - e_k(\hat{\omega}) = M_k - e_k(\omega) - (e_k(\hat{\omega}) - e_k(\omega)).$$

For any $a \in R^K$, denote $T_n a = a_T = (a_{T,1}, \cdots, a_{T,K})'$. Then,

$$\begin{aligned} a'(n^{-1/2}T_n'\xi_n) &= n^{-1/2}a_T'\xi_n \\ &= n^{-1/2}\textstyle\sum_{k=1}^{K} a_{T,k}(M_k - e_k(\omega)) - n^{-1/2}\sum_{k=1}^{K} a_{T,k}(e_k(\hat{\omega}) - e_k(\omega)). \end{aligned} \tag{27}$$

Denote $p_\omega(y_i = 1) = \pi_i(\omega)$. By Taylor expansion and Theorem 1, we have

$$a'(n^{-1/2}T_n'\xi_n) = \sum_{i=1}^{n}\Upsilon_i = n^{-1/2}\sum_{i=1}^{n}a_n'h_i^{(1)},$$

where $\Upsilon_i = n^{-1/2}\lambda_n' h_i^{(1)}$ is an array of martingale differences. The remaining proof is omitted because it is similar to that in Jiang (2001a). Only difference is that the asymptotics here will be proved by using a martingale central limit theorem. Because of the use of partial likelihood, this result is more general than Jiang (2001a).

Lemma D.2. *Using the notation in Section 3.3, for any* $\mu \in R \setminus \{0\}$,

$$\mu J^{-1/2} I^N \sqrt{m}(\hat{\sigma}_b^2 - \sigma_b^2) = [(\mathbf{Y} - \mathbf{X}\omega)' B_N^* (\mathbf{Y} - \mathbf{X}\omega) - E((\mathbf{Y} - \mathbf{X}\omega)' B_N^* (\mathbf{Y} - \mathbf{X}\omega))], \tag{28}$$

where $B_N^* = J^{-1/2}\mu \mathbf{W}^{1/2} V^* \mathbf{W}^{1/2} Z Z' \mathbf{W}^{1/2} V^* \mathbf{W}^{1/2}/\sqrt{m}$.

Proof:

Follow the same argument as in Theorem 2, consider the LMM with GLM weights, we can obtain this result by modifying the first formula on page 276 of Jiang (1996) into

$$\mu J^{-1/2} I^N \sqrt{m}(\hat{\sigma}_b^2 - \sigma_b^2) = \varpi' B_N \varpi - \mathrm{E}(\varpi' B_N \varpi),$$

where $B_N = J^{-1/2}\mu V_1(\sigma_b)/\sqrt{m}$, and $V_1(\sigma_b)$ is defined in Theorem 2. Lemma D.2 follows because $\varpi' g(\sigma_b)\mathbf{W}^{-1/2} = (\mathbf{Y} - \mathbf{X}\omega)'$.

Lemma D.3. *Denote* $\theta = (\omega', \sigma_b)$, $\xi_k = M_k - e_k(\hat{\theta})$, *and* $\xi = (\xi_k)_{1 \le k \le K}$. *Let* T *be an orthogonal matrix such that* $T'\Psi_N T = diag(\lambda_{N,1}, \cdots, \lambda_{N,K})$, *where* $\lambda_{N,1}, \cdots, \lambda_{N,K}$ *are the eigenvalues of* Ψ_N. *For any* $a \in R^K$,

$$a'((N)^{-1/2} T'\xi) = \sum_{i=1}^{m}\sum_{t=1}^{n} \Upsilon_{it} + o_p(1), \tag{29}$$

where $Ta = a_T = (a_{T,1}, \cdots, a_{T,K})'$, $G_{it}^* = (N)^{-1/2} a_T' G_{it}$, $\Delta' = (\Delta_{it}) = (\mathbf{Y} - \mathbf{X}\omega)'\mathbf{V}^{-1/2}$, $Var(\Delta_{it}) = 1$, *and*

$$\begin{aligned} \Upsilon_{it} &= G_{it}^*(y_{it} - \pi_{it}(\theta)) - (C^*\mathbf{V})_{it}\Delta_{it}^2 \\ &\quad -(\textstyle\sum_{(it)' \neq it}(C^*\mathbf{V})_{it,(it)'}\Delta_{(it)'})\Delta_{it} + (C^*\mathbf{V})_{it}. \end{aligned}$$

(Notice that C^* is a $N \times N$ matrix with $C_{i't',it}^*$ indicating the element in C^* with

$[(i'-1)n+t']$-th column and $(i-1)n+t$-th row, and by the definition before Theorem 3, $C^*_{it,it} = C^*_{it}$.)

Proof:

For $1 \le k \le K$,

$$\xi_k = M_k - e_k(\theta) - (e_k(\hat{\theta}) - e_k(\theta)).$$

By definition,

$$\begin{aligned} a'((N)^{-1/2}T'\xi) = & \ (N)^{-1/2}a_T'\xi \\ = & \ (N)^{-1/2}\textstyle\sum_{k=1}^K a_{T,k}(M_k - e_k(\theta)) - (N)^{-1/2}\sum_{k=1}^K a_{T,k}(e_k(\hat{\theta}) - e_k(\theta)). \end{aligned}$$

For the first term on the right hand side, denoting $p_\theta(y_i = 1) = \pi_i(\theta)$, we have

$$(N)^{-1/2}\sum_{k=1}^K a_{T,k}(M_k - e_k(\theta)) = (N)^{-1/2}\sum_{i=1}^m\sum_{t=1}^n a_T'\left[\mathbf{1}_{(H_i \in E_k)}\right]_{1\le k\le K}(y_{it} - \pi_{it}(\theta)).$$

For the second term, by Taylor expansion, we have

$$(N)^{-1/2}\sum_{k=1}^K a_{T,k}(e_k(\hat{\theta}) - e_k(\theta)) = (N)^{-1/2}\sum_{k=1}^K a_{T,k}\left[\sum_{i=1}^m\sum_{t=1}^{n_i}\left(\mathbf{1}_{(H_{it}\in E_k)}\left(\pi_{it}(\hat{\theta}) - \pi_{it}(\theta)\right)\right)\right] \tag{30}$$

$$\approx (N)^{-1/2}\sum_{k=1}^K a_{T,k}\left[\left(\sum_{i=1}^m\sum_{t=1}^n \mathbf{1}_{(H_{it}\in E_k)}\frac{\partial}{\partial\omega'}\pi_{it}(\theta)\right)(\hat{\omega} - \omega) + \left(\sum_{i=1}^m\sum_{t=1}^n \mathbf{1}_{(H_{it}\in E_k)}\frac{\partial}{\partial\sigma_b^2}\pi_{it}(\theta)\right)(\hat{\sigma}_b^2 - \sigma_b^2)\right].$$

Assume $D^* = (N)^{-1/2}\sum_k a_{T,k}\left(\frac{1}{N}\sum_i\sum_t \mathbf{1}_{(H_{it}\in E_k)}\frac{\partial}{\partial\omega'}\pi_{it}(\theta)\right)\Lambda_N^{-1} = N^{-1/2}a_T'D$.

By Theorem 1, Assumption A5, and Lemma D.2 with

$$\mu = (N)^{-1/2}a_T'\left[\sum_i\sum_t(\mathbf{1}_{(H_{it}\in E_k)}\frac{\partial}{\partial\sigma_b^2}\pi_{it}(\theta))\right]\frac{(J^{-1/2}I^N)^{-1}}{\sqrt{m}},$$

we have

$$\begin{aligned} & (N)^{-1/2}\textstyle\sum_k a_{T,k}(e_k(\hat{\theta}) - e_k(\theta)) \\ & = D^*\mathbf{X}'(\mathbf{y} - \pi(\theta)) + (\mathbf{Y} - \mathbf{X}\omega)'C^*(\mathbf{Y} - \mathbf{X}\omega) - \mathrm{E}((\mathbf{Y} - \mathbf{X}\omega)'C^*(\mathbf{Y} - \mathbf{X}\omega)) + o_p(1) \\ & = \textstyle\sum_{i=1}^m\sum_{t=1}^n\Big\{D^*X_{it}(y_{it} - \pi_{it}(\theta)) + (C^*\mathbf{V}_{it})\Delta_{it}^2 \\ & + (\textstyle\sum_{(it)'\neq it}(C^*\mathbf{V})_{it,(it)'}\Delta_{(it)'})\Delta_{it} - (C^*\mathbf{V})_{it}\Big\} + o_p(1), \end{aligned}$$

where $C^* = (N)^{-1/2} a_T' \left[\sum_i \sum_t (\mathbf{1}_{(H_{it} \in E_k)} \frac{\partial}{\partial \sigma_b^2} \pi_{it}(\theta)) \right]_{1 \le k \le K} \frac{(I^N)^{-1}}{\sqrt{m}} C = (N)^{-1/2} a_T' \Phi C.$

Lemma D.4. *Under A6-A8, as $N \to \infty$,*

$$\sum_{i=1}^{m} \sum_{t=1}^{n} \Upsilon_{it} \to_d \mathcal{N}(0, a' \Gamma a). \tag{31}$$

Proof:

First, derive the asymptotic distribution of $\sum_{i=1}^m \sum_{t=1}^n \Upsilon_{it}/\sigma_N$, where σ_N is defined in A6. Decompose $\Upsilon_{it}/\sigma_N = \Upsilon_{it}^{(1)} + \Upsilon_{it}^{(2)}$, where

$$\Upsilon_{it}^{(1)} = \frac{1}{\sigma_N} \left(G_{it}^* u_{it}^* + (C^* \mathbf{V})_{it} U_{it} + \sum_{(it)' \neq it} \left((C^* \mathbf{V})_{it,(it)'} u_{(it)'} \right) u_{it} \right),$$

$$\Upsilon_{it}^{(2)} = \frac{1}{\sigma_N} \left(G_{it}^* v_{it}^* + (C^* \mathbf{V})_{it} V_{it} + \left(\sum_{(it)' \neq it} (C^* \mathbf{V})_{it,(it)'} v_{(it)'} \right) u_{it} + \left(\sum_{(it)' \neq it} \Delta_{(it)'} (C^* \mathbf{V})_{it,(it)'} \right) v_{it} \right).$$

Define

$$\begin{aligned}
U_{it} &= (\Delta_{it}^2 - 1) \mathbf{1}_{(|\Delta_{it}| < L_{it})} - \mathrm{E}(\Delta_{it}^2 - 1) \mathbf{1}_{(|\Delta_{it}| < L_{it})}, \\
V_{it} &= (\Delta_{it}^2 - 1) - U_{it}, \\
u_{it} &= \Delta_{it} \mathbf{1}_{(|\Delta_{it}| < L_{it})} - \mathrm{E} \Delta_{it} \mathbf{1}_{(|\Delta_{it}| < L_{it})}, \\
v_{it} &= \Delta_{it} - u_{it}, \\
u_{it}^* &= (y_{it} - \pi_{it}(\theta)) \mathbf{1}_{(|y_{it} - \pi_{it}(\theta)| < L_{it})} - \mathrm{E}(y_{it} - \pi_{it}(\theta)) \mathbf{1}_{(|y_{it} - \pi_{it}(\theta)| < L_{it})}, \\
v_{it}^* &= (y_{it} - \pi_{it}(\theta)) - u_{it}^*.
\end{aligned}$$

By assumption A6, we can easily show that $\sum_{i=1}^m \sum_{t=1}^n \Upsilon_{it}^{(2)}$ converge to 0 in L_2. Next, consider $\Upsilon_{it}^{(1)}$ which is an array of martingale differences by following the same argument in Theorem 5.2 of Jiang (1996). Based on assumption A7 and Rosenthal's inequality (Hall and Heyde, 1980), $\max_{it} \mid \Upsilon_{it}^{(1)} \mid$ is bounded in L_2 and converges to 0 in probability.

By Theorem 3.2 of Hall and Heyde (1980), to prove Lemma D.4, one has to show

that $\sum_i \sum_t (\Upsilon_{it}^{(1)})^2$ converge to $a'\Gamma a$ in probabilty. First, it can be decomposed as

$$\sum_{i=1}^{m}\sum_{t=1}^{n}(\Upsilon_{it}^{(1)})^2 = \sum_{j=1}^{3} t_i + \sum_{j=1}^{3} s_i,$$

where

$$\begin{aligned}
t_1 &= \sigma_N^{-2}\textstyle\sum_{i=1}^{m}\sum_{t=1}^{n}[((C^*\mathbf{V})_{it}U_{it} + G_{it}^*u_{it}^*)^2 - \mathrm{E}((C^*\mathbf{V})_{it}U_{it} + G_{it}^*u_{it}^*)^2],\\
t_2 &= 2\sigma_N^{-2}\textstyle\sum_{i=1}^{m}\sum_{t=1}^{n}(\sum_{(it)'\neq it}(C^*\mathbf{V})_{it,(it)'}u_{(it)'})[(C^*\mathbf{V})_{it}(U_{it}u_{it} - \mathrm{E}(U_{it}u_{it}))\\
&\quad +(G_{it}^*u_{it}^*)u_{it} - \mathrm{E}((G_{it}^*u_{it}^*)u_{it})],\\
t_3 &= 2\sigma_N^{-2}\textstyle\sum_{i=1}^{m}\sum_{t=1}^{n}\left((\sum_{(it)'\neq it}(C^*\mathbf{V})_{it,(it)'}u_{(it)'})^2(u_{it}^2 - \mathrm{E}u_{it}^2)\right),\\
s_1 &= \sigma_N^{-2}\textstyle\sum_{i=1}^{m}\sum_{t=1}^{n}\mathrm{E}((C^*\mathbf{V})_{it}U_{it} + G_{it}^*u_{it}^*)^2,\\
s_2 &= 2\sigma_N^{-2}\textstyle\sum_{i=1}^{m}\sum_{t=1}^{n}(\sum_{(it)'\neq it}(C^*\mathbf{V})_{it,(it)'}u_{(it)'})[\mathrm{E}((C^*\mathbf{V})_{it}U_{it}u_{it}) + \mathrm{E}((G_{it}^*u_{it}^*)u_{it})],\\
s_3 &= 2\sigma_N^{-2}\textstyle\sum_{i=1}^{m}\sum_{t=1}^{n}\left((\sum_{(it)'\neq it}(C^*\mathbf{V})_{it,(it)'}u_{(it)'})^2\mathrm{E}u_{it}^2\right).
\end{aligned}$$

By assumption A7 and Rosenthal's inequality, we can show that $t_i \to 0$ in L_2 for $i = 1, 2, 3$, which is similar to the result in Theorem 5.2 of Jiang (1996). By assumption A7,

$$s_1 = \sigma_N^{-2}\sum_{i=1}^{m}\sum_{t=1}^{n}\mathrm{Var}\left((C^*\mathbf{V})_{it}\Delta_{it}^2 + G_{it}^*(y_{it} - \pi_{it}(\theta))\right) + o_p(1). \tag{32}$$

Analogue to Theorem 5.2 of Jiang (1996), by assumptions A6-A8, we have

$$Es_2^2 \le c\left[\frac{\lambda\max(\xi'\xi)}{\sigma_N^2}\right]^{1/2} \to 0, \tag{33}$$

where c stands for a constant and

$$s_3 = 2\sigma_N^{-2}\sum_{(it)'\neq it}(C^*\mathbf{V})_{it,(it)'}^2 + o_p(1). \tag{34}$$

By (32) and (34), $\sum_i \sum_t (\Upsilon_{it}^{(1)})^2 = 1 + o_p(1)$. Because $\sigma_N^2 = a'T'\Psi_N Ta$, it converges to $a'\Gamma a$ in probability. Consequently, (31) follows.

Proof of Theorem 3:

From Lemmas D.2 to D.4, we have, for any a,

$$a'(N^{-1/2}T'\xi) \to_d (a'\Gamma a)^{1/2}\mathbb{Z},$$

where $\mathbb{Z} \sim \mathcal{N}(0, \mathbf{1})$, form which, $N^{-1/2}T'\xi \rightarrow_d \mathcal{N}(0, \Gamma)$ follows.

REFERENCES

Albert, A. and Anderson, J. A. (1984), "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika* **71**, 1-10.

Barndorff-Nielsen, O. E. and Cox, D. R. (1989), *Asymptotic Techniques for Use in Statistics,* London: Chapman and Hall.

Benjamin, M. A., Robert, A. R., and Stasinopoulos, D. M. (2003), "Generalized Autoregressive Moving Average Models," *Journal of American Statistical Association*, **98**, 214-223.

Breiman, L. (1995), "Better Subset Regression using the Nonnegative Garrote," *Technometrics*, **37**, 373-384.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate Inference to Generalized Linear Mixed Models," *Journal of American Statistical Association* **88**, 9-25.

Cappelleri, D. J., Frecker, M. I., Simpson, T. W., and Snyder, A. (2002), "Design of a PZT Bimorph Actuator Using a Metamodel-Based Approach," *ASME Journal of Mechanical Design*, **124**, 354-357.

Chen, W., Jin, R., and Sudjianto, A. (2005), "Analytical Variance-Based Global Sensitivity Analysis in Simulation-Based Design Under Uncertainty," *ASME Journal of Mechanical Design*, **127**, 875-886.

Chernoff, H. and Lehmann, E. L. (1954), "The Use of Maximum Likelihood Estimations in χ^2 Tests for Goodness of fit," *The Annals of Mathematical Statistics* **25**, 579-586.

Chesla, S. E., Selvaraj, P., and Zhu, C. (1998), "Measuring Two-Dimensional Receptor-Ligand Binding Kinetics by Micropipette," *Biophysical Journal* **75**, 1553-1572.

Chipman, H., Hamada, M. and Wu, C. F. J. (1997), "A Bayesian Variable Selection Approach for Analyzing Designed Experiments with Complex Aliasing," *Technometrics*, **39**, 372-381.

Cox, D. R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society,* Ser. B, **34**, 187-202.

Cox, D. R. (1975), "Partial Likelihood," *Biometrika* **62**, 69-76.

Currin, C., Mitchell, T. J., Morris, M. D. and Ylvisaker, D. (1991), "Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments," *Journal of American Statistical Association*, **86**, 953-963.

Diggle, P., Heagerty, P., Liang, K-Y., and Zeger, S. (2002), *Analysis of Longitudinal Data,* 2nd Ed. Oxford: Oxford University Press.

Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics*, **32**, 407-499.

Fang, K. T., Ma, C. X. and Winker, P. (2002), "Centered L_2-discrepancy of random sampling and Latin hypercube design, and construction of uniform designs," *Mathematics of Computation*, **71**, 275-296.

Fang, K. T., Li, R., and Sudjianto, A. (2006), *Design and Modeling for Computer Experiments,* CRC Press, New York.

Fokianos, K. and Kedem, B. (2004), "Partial Likelihood Inference for Time Series Following Generalized Linear Models," *Journal of Time Series Analysis* **25**, 173-197.

George, E. I. and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of American Statistical Association*, **88**, 881-889.

Hall, P. and Heyde, C. C. (1980), *Martingale Limit Theory and its Application,* New York: Academic Press.

Hamada, M. and Wu, C. F. J. (1992), "Analysis of Designed Experiments with Complex Aliasing," *Journal of Quality Technology*, **24**, 130-137.

Harville, D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," *Journal of American Statistical Association* **72**, 320-340.

Hicks, C. R. and Turner K. V. (1999), *Fundamental Concepts in the Design of Experiments,* 5th ed., Oxford: Oxford University Press, New York.

Hoffman, R. M., Sudjianto, A., Du, X., Stout, J. (2003), "Robust Piston Design and Optimization Using Piston Secondary Motion Analysis," SAE Paper 2003-01-0148, *SAE Transactions.*

Iman, R. L. and Conover, W. J. (1982), "A distribution-free approach to inducing rank correlation among input variables," *Communication and Statistics, Part B–Simulation and Computation*, **11**, 311-334.

Jiang, J. (1996), "REML Estimation: Asymptotic Behavior and Related Topics," *The Annals of Statistics* **24**, 255-286.

Jiang, J. (2001a), "A Nonstandard Chi-square Test with Application to Generalized Linear Model Diagnostics," *Statistics and Probability Letters* **53**, 101-109.

Jiang, J. (2001b), "Goodness-of-fit Tests for Mixed Model Diagnostics," *The Annals of Statistics* **29**, 1137-1164.

Jin, R., Chen, W., and Simpson, T. (2001), "Comparative Studies of Metamodeling Techniques under Multiple Modeling Criteria," *Journal of Structural & Multidisciplinary Optimization*, **23**, 1-13.

Jin, R., Chen, W. and Sudjianto, A. (2005), "An efficient algorithm for constructing optimal design of computer experiments," *Journal of Statistical Planning and Inference*, **134**, 268-287.

Johnson, M., Moore, L. and Ylvisaker, D. (1990), "Minimax and maximin distance design," *Journal of Statistical Planning and Inference*, **26**, 131-148.

Joseph, V. R. (2006a), "Limit kriging," *Technometrics*, **48**, 458-466.

Joseph, V. R. (2006b), "A Bayesian Approach to the Design and Analysis of Fractionated Experiments," *Technometrics*, **48**, 219-229.

Joseph, V. R. and Delaney, J. D. (2007), "Functionally Induced Priors for the Analysis of Experiments," *Technometrics*, **49**, 1-11.

Joseph, V. R. and Hung, Y. (2008), "Orthogonal-maximin Latin Hypercube Designs," *Statistica Sinica*, **18**, 171-186.

Kaufmann, H. (1987), "Regression Models for Nonstationary Categorical Time Series: Asymptotic Estimation Theory," *The Annals of Statistics* **15**, 79-98.

Kedem, B. and Fokianos, K. (2002), *Regression Models for Time Series Analysis,* New York: Wiley.

Li, R. and Sudjianto, A. (2005), "Analysis of Computer Experiments Using Penalized Likelihood in Gaussian Kriging Models," *Technometrics*, **47**, 111-120.

Li, W, K. (1994), "Time Series Models Based on Generalized Linear Models: Some Further Results," *Biometrics*, **50**, 506-511.

Li, W. W. and Wu, C. F. J. (1997), "Columnwise-pairwise algorithms with applications to the construction of supersaturated designs," *Technometrics*, **39**, 171-179.

Lin, X. and Breslow, N. E. (1996), "Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion," *Journal of American Statistical Association* **91**, 1007-1016.

Lundy, M. and Mees, A. (1986), "Convergence of an annealing algorithm," *Mathematical Programming*, **34**, 111-124.

Marshall, B. T., Long, M., Piper, J. W., Yago, T., McEver, R. P., and Zhu, C. (2003), "Direct Observation of Catch Bonds Involving Cell-adhesion Molecules," *Nature* **423** 190-193.

Martin, J. D. and Simpson, T. W. (2005), "On the Use of Kriging Models to Approximate Deterministic Computer Models," *AIAA Journal*, **43**, 853-863.

Marusich, T. D. and Ortiz, M. (1995), "Modeling and Simulation of High-Speed Machining," *International Journal for Numerical Methods in Engineering*, **38**, 3675-3694.

Mehta, A. D., Rief, M., Spudich, J. A., Smith, D. A., and Simmons, R. M. (1999), "Single-molecule Biomechanics with Optical Methods." *Science* **283**(5408) 1689-1695.

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of American Statistical Association* **92**, 162-170.

McKay, M. D., Beckman, R. J. and Conover, W. J. (1979), "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, **21**, 239-245.

Miller, A. (2002), *Subset Selection in Regression*. CRC Press, New York.

Montgomery, D. C. (2004), *Design and Analysis of Experiments*, Wiley, New York.

Morris, M. D. and Mitchell, T. J. (1995), "Exploratory Designs for Computer Experiments," *Journal of Statistical Planning and Inference*, **43**, 381-402.

Morris, M. D., Mitchell, T. J., and Ylvisaker, D. (1993), "Bayesian Design and Analysis of Computer Experiments: Use of Derivatives in Surface Prediction," *Technometrics*, **35**, 243-255.

Owen, A. (1994), "Controlling correlations in Latin hypercube samples," *Journal of American Statistical Association*, **89**, 1517-1522.

Pacheco, J. E., Amon, C. H., and Finger, S. (2003), "Bayesian Surrogates Applied to Conceptual Stages of the Engineering Design Process," *ASME Journal of Mechanical Design*, **125**, 664-672.

Pan, W. (2001), "On the Robust Variance Estimator in Generalised Estimating Equations," *Biometrika* **88**, 901-906.

Park, J. S. (1994), "Optimal Latin-hypercube designs for computer experiments," *Journal of Statistical Planning and Inference*, **39**, 95-111.

Patterson, H. D. and Thompson, R. (1971), "Recovery of Interblock Information When Block Sizes Are Unequal," *Biometrika* **58**, 545-554.

Phadke, M. S. (1989), *Quality Engineering Using Robust Design*, Prentice Hall, Englewood Cliffs, NJ.

Qian, Z., Seepersad, C. C., Joseph, V. R., Allen, J. K. and Wu, C. F. J (2006), "Building surrogate models based on detailed and approximate simulations," *ASME Journal of Mechanical Design*, **128**, 668-677.

Sacks, J., Schiller, S. B., and Welch, W. J. (1989), "Design of Computer Experiments," *Technometrics*, **31**, 41-47.

Sacks, J., Welch, W. J., Mitchell, T. J. and Wynn, H. P. (1989), "Design and analysis of computer experiments," *Statistical Science*, **4**, 409-423.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments,* Springer, New York.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, New York: Wiley.

Senatov, V. V. (1980), "Uniform Estimates of the Rate of Convergence in the Multidimensional Central Limit Theorem," *Theory of Probability and its Applications* **25**, 745-759.

Shewry, M. C. and Wynn, H. P. (1987), "Maximum entropy sampling," *Journal of Applied Statistics*, **14**, 165-170.

Silvapulle, M. (1981), " On the Existence of Maximum Likelihood Estimates for the Binomial Response Models," *Journal of the Royal Statistical Society,* Ser. B, **43**, 310-313.

Slud, E. (1992), "Partial Likelihood for Continuous-time Stochastic Processes," *Scandinavian Journal of Statistics* **19**, 97-109.

Slud, E. and Kedem, B. (1994), "Partial Likelihood Analysis of Logistic Regression and Autoregression," *Statistica Sinica*, **4**, 89-106.

Taguchi, G. (1987), *Systems of Experimental Design, Vol.1 & Vol 2*, White Plains, New York: Unipub/Kraus International.

Tang, B. (1993), "Orthogonal array-based Latin hypercubes," *Journal of American Statistical Association*, **88**, 1392-1397.

Tang, B. (1998), "Selecting Latin hypercubes using correlation criteria," *Statistica Sinica*, **8**, 965-978.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B*, **58**, 267-288.

Tierney, L. and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of American Statistical Association* **81**, 82-86.

Wackernagel, H. (2002), *Multivariate Geostatistics*, Springer, New York.

Wedderburn, R. W. M. (1976), "On the Existence and Uniqueness of the Maximum Likelihood Estimates," *Biometrika* **63**, 27-32.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., and Morris, M. D. (1992), "Screening, Predicting, and Computer Experiments," *Technometrics*, **34**, 15-25.

Winker, P. and Fang, K. T. (1998), "Optimal U-type design," in *Monte Carlo and Quasi-Monte Carlo Methods* (1996), eds. H. Niederreiter, P. Zinterhof and P. Hellekalek, Springer, 436-448.

Wong, W. H. (1986), "Theory of Partial Likelihood," *The Annals of Statistics* **14**, 88-123.

Wu, C. F. J., and Hamada, M. (2000), *Experiments: Planning, Analysis, and Parameter Design Optimization,* Wiley, New York.

Ye, K.Q. (1998), "Orthogonal column Latin hypercubes and their application in computer experiments," *Journal of American Statistical Association*, **93**, 1430-1439.

Ye, K. Q., Li, W. and Sudjianto, A. (2000), "Algorithmic construction of optimal symmetric Latin hypercube designs," *Journal of Statistical Planning and Inference*, **90**, 145-159.

Zarnitsyna, V. I., Huang, J., Zhang, F., Chien, Y-H., Leckband, D., and Zhu, C. (2007), "Memory in Receptor-ligand Mediated Cell Adhesion," *Proceedings of the National Academy of Sciences USA.*, **104**, 18037-18042.

Zeger, S. L. and Qaqish, B. (1988), "Markov Models for Time Series: A Quasi-Likelihood Approach," *Biometrics*, **44**, 1019-1032.

Zhu, C., Long, M., Chesla, S. E., and Bongrand, P. (2002), "Measuring Receptor/Ligand Interaction at the Single-bond Level: Experimental and Interpretative Issues," *Annals of Biomedical Engineering* **30**, 305-314.